Study Note 98-04

Application of Lightfoot's Cluster Evaluation System to Current Problems in Army Occupational Analysis

Mary Ann Lightfoot
Human Resources Research Organization

Tirso E. Diaz
Human Resources Research Organization

Yefim Vladimirsky
Human Resources Research Organization

January 1998

United States Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Directorate of the U.S. Total Army Personnel Command

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Peter M. Greenston

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to : U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: TAPC-ARI-PO, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1998, January	Final	September 1995-February 1997

4. TITLE AND SUBTITLE Application of Lightfoot's Cluster Evaluation System to Current Problems in Army Occupational Analysis	5a. CONTRACT OR GRANT NUMBER MDA903-93-D-0032 DO 0036
	5b. PROGRAM ELEMENT NUMBER 433709
6. AUTHOR(S) Mary Ann Lightfoot, Tirso E. Diaz, and Yefim Vladimirsky	5c. PROJECT NUMBER
	5d. TASK NUMBER 7002
	5e. WORK UNIT NUMBER C02
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization (HumRRO) 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314	8. PERFORMING ORGANIZATION REPORT NUMBER FR-WATSD-97-04
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-RS 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Study Note 98-04

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

COR: Michael G. Rumsey

14. ABSTRACT *(Maximum 200 words)*:

The study objective was to build a prototype cluster structure validation methodology and to test it in a population data base of Army military occupational specialties. We developed a cross-validation and internal validity (CV*IV) procedure for estimating the cluster structures of empirical data bases. The major contributions of the CV*IV procedure are that it can be used with many different types of empirical data and includes a statistical approach for identifying optimal cluster structure. We validated the CV*IV procedure through an experimental design that allowed us to analyze the properties of the statistical test in terms of Type I error rate, power, and precision. The results provide strong support for the validity and utility of the CV*IV procedure for estimating population cluster structure from sample data. First, the statistical test preserved the Type I error rate of .05. Second, the power of the test ranged between 86% and 100% across sample sizes. Third, 63% of the sample results matched the cluster structure of the Army job population data base. The CV*IV procedure has wide application for the analysis of cluster structures in a range of data bases in both research and applied settings across the social and physical sciences.

15. SUBJECT TERMS

Cluster analysis   Occupational classification   Job families
Job analysis   Occupational analysis

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	101	

Study Note 98-04

Application of Lightfoot's Cluster Evaluation System to Current Problems in Army Occupational Analysis

Mary Ann Lightfoot
Human Resources Research Organization

Tirso E. Diaz
Human Resources Research Organization

Yefim Vladimirsky
Human Resources Research Organization

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

January 1998

Army Project Number
20433709

Army Personnel Management
and Support Activity

FOREWORD

The diverse actual and potential uses of Army Occupational Analysis (OA) Program data throughout all phases of the Personnel Manning Life Cycle of the Army make investigating innovative approaches for expanding the automated data analysis capacity of the OA program a key goal. At the heart of this effort is the use of reliable and valid procedures to cluster tasks and other types of job data into meaningful units. Occupational clustering is essentially a data reduction strategy designed to improve our understanding of the structure of work by reducing the complexity of job information and finding the underlying patterns. Further, clustering is a tool for simplifying manpower, personnel, and training procedures.

Many software packages have been developed to address this need. However, there are no well established, accepted methods for evaluating the reliability and validity of the occupational cluster structures produced by the various clustering techniques, or of occupational groupings developed by subject matter experts. This gap in analytical procedures limits the quality and utility of occupational analysis products at all levels, from the grouping of tasks into task modules through the creation, combination, or elimination of Military Occupational Specialties (MOS), up to the grouping of MOS into Career Management Fields.

The objective of this project was to evaluate the scientific properties and practical utility for meeting Army occupational analysis needs of a cluster reliability and validation method (CRVM) developed in an earlier project, entitled *Occupational Analysis and Job Structures*. This report describes a statistical application of the CRVM, the cluster structure cross-validation and internal validity (CV*IV) procedure. It also presents the results of an evaluation of the technique using job analysis data from the population of Army MOS. Although the CRVM and CV*IV procedure were developed for industrial and organizational psychology research and applications, they are also useful tools for other social and physical science disciplines concerned with creating reliable and valid classification structures.

ZITA M. SIMUTIS
Technical Director

ACKNOWLEDGMENTS

This report is dedicated to the memory of Dr. Dickie A. Harris, who provided strong support and valuable suggestions during the course of this and an earlier project which led to the development and evaluation of a new statistical technique for measuring the reliability and validity of cluster structures. The authors would like to thank Dr. Paul Sticha, who made important contributions to the early stages of this project through his suggestions concerning conceptualization of the technique and by developing imaginative solutions to several problems that arose during the course of the project. We would like to thank Dr. Clinton B. Walker, who saw both the practical and scientific potentials of the cluster reliability and validation method early on and provided enthusiastic support throughout the test and evaluation process. We would also like to thank Mr. Darrel A. Worstine, Chief of ARI's Occupational Analysis Office, for his early guidance on how to tailor a cluster validation procedure to be most useful to Army occupational analysts, and for his interest in testing it on current Army data bases. Finally, we want to express our special thanks to and appreciation of all our colleagues at HumRRO who have shown so much interest in and support of our efforts, especially Dr. Brian A. Waters, Dr. Janice H. Laurence, and Ms. Jennifer A. Naughton.

APPLICATION OF LIGHTFOOT'S CLUSTER EVALUATION SYSTEM TO CURRENT PROBLEMS IN ARMY OCCUPATIONAL ANALYSIS

EXECUTIVE SUMMARY

Research Requirement:

The objective of this study was to build a prototype cluster structure validation methodology and to test it in a population data base of Army military occupational specialties (MOS).

Procedure:

Based on preliminary research and development conducted in an earlier study by Statman, Gribben, Harris and Hoffman (1994) entitled, *Occupational Analysis and Job Structures*, we developed a prototype cross-validation and internal validity (CV*IV) procedure for estimating the cluster structures of empirical data bases. The two major contributions of the CV*IV procedure are that it can be used with many different types of empirical data and include a statistical approach for identifying optimal cluster structure. Most previous cluster validation research has been conducted on synthetic data and has limited relevance to real data. Further, most cluster analysis techniques do not include statistical procedures and, therefore, are confined to exploratory, rather than inferential, data analysis.

We validated the CV*IV procedure through an experimental design that allowed us to analyze the properties of the statistical test in terms of Type I error rate, power, and precision. We used a Monte Carlo procedure to create 100 random samples to study the actual Type I error rate of the CV*IV procedure. We also conducted 300 replications of the experiment through repeated sampling from an empirical population data base of Army jobs to test the power and precision of the procedure. We varied our analyses to reflect small, moderate, and large samples and two distributions under the null hypothesis.

Findings:

The results provide strong support for the validity and utility of the CV*IV procedure for estimating population cluster structure from sample data. First, the statistical test preserved the Type I error rate of .05. Second, the power of the test ranged between 86% and 100% across sample sizes. Third, 63% of the sample results matched the cluster structure of the Army job population data base. Fourth, the distribution of the null population did not affect the results of the CV*IV procedure.

Utilization of Findings:

The CV*IV procedure has wide application for the analysis of cluster structures in a range of data bases in both research and applied settings across the social and physical sciences. The CV*IV procedure should be especially useful to the Army and other Services for analyzing the military occupational structure in the present environment of changing missions and rapid advances in computer and telecommunications technology.

APPLICATION OF LIGHTFOOT'S CLUSTER EVALUATION SYSTEM TO CURRENT PROBLEMS IN ARMY OCCUPATIONAL ANALYSIS

CONTENTS

Page

Appendixes

List of Tables and Figures

APPLICATION OF LIGHTFOOT'S CLUSTER EVALUATION SYSTEM TO CURRENT PROBLEMS IN ARMY OCCUPATIONAL ANALYSIS

Introduction

We describe in this report a three-stage cluster reliability and validation method (CRVM) and an application of the first two stages of the CRVM, called the cluster structure cross-validation and internal validation (CV*IV) procedure. The contribution of the CRVM to cluster analysis and occupational classification is that it provides a systematic approach for measuring the accuracy of cluster structures in real (i.e., empirical not synthetic) data. The CV*IV procedure is a refinement of the CRVM in that it adds a statistical hypothesis test to the first two components of the method.

This report is divided into four chapters. In chapter one, Introduction, we discuss the set of problems that led to the proposal of the CRVM and present an overview of the approach. In chapter two, Method, we describe the CV*IV procedure. Chapter three, Experimental Validation of the CV*IV procedure, presents the test and evaluation of the CV*IV procedure, including both experimental design and results. The last chapter, Discussion and Conclusions, presents a summary of the results, a discussion of the limitations of the CV*IV procedure and suggestions for future research. Appendix E, Suggested Applications of the CV*IV Procedure for Determining the Population Cluster Structure, contains four examples of how to use the CV*IV procedure and presents outputs which illustrate a range of possible results.

The Need for Methods to Evaluate the Accuracy of Cluster Structures

The Study

This study is the second part of a two-part project. The initial research and development is described in Statman[1], Gribben, Harris, and Hoffman (1994) and Statman (1996). The catalyst for the project was a question, which had both theoretical and practical ramifications for occupational analysis and classification, posed by researchers at the Army Research Institute (ARI): What is the occupational structure of Army jobs, known as military occupational specialties (MOS)? This question is important for Army researchers and occupational analysts who develop personnel systems and conduct other research to improve the readiness of the force and the effectiveness and efficiency of the manpower, personnel and training systems that support the Army's peacetime and wartime missions.

ARI researchers expressed three specific concerns underlying their question about the structure of Army occupations. First, the Army has two separate classification systems for grouping MOS into larger categories: the Aptitude Area (AA) structure and the Career Management Fields (CMF). These occupational groupings contain different numbers of MOS clusters (9 for the AAs and 35 for the CMFs) and are not entirely consistent. Tables 1 and 2 present the two occupational structures. Appendix A lists the MOS investigated in this project

---

[1] The first author recently changed her name to Lightfoot and is the first author of the present report.

and their AA and CMF classifications. The AA structure groups MOS according to similarities in ability requirements for selection of recruits into the Army and assignment into specific jobs. The CMF group jobs into career ladders that guide training and promotion decisions. Statman et al. (1994) describe the development of the AA and CMF structures. The first question posed by ARI's researchers was whether the differences in the two occupational structures were valid, reflecting real distinctions in their operational uses.

Table 1. Army Aptitude Areas

• Combat	Infantry, Armor, Combat Engineer
• Field Artillery	Field Cannon and Rocket Artillery
• Electronics Repair	Missiles Repair, Air Defense Repair, Tactical Electronic Repair, Fixed Plant Communications Repair
• Operators and Food	Missiles Crewman, Air Defense Crewman, Driver, Food Services
• Surveillance and Communications	Target Acquisition and Combat Surveillance, Communication Operations
• Mechanical Maintenance	Mechanical and Air Maintenance, Rails
• General Maintenance	Construction and Utilities, Chemical, Marine, Petroleum
• Clerical	Administrative, Finance, Supply
• Skilled Technical	Medical, Military Policeman, Intelligence, Data Processing, Air Control, Topography and Printing, Information and Audio Visual

Source: Maier & Fuchs, 1972.

Table 2. Enlisted Career Management Fields

- Administration
- Recruiting and Reenlistment
- Public Affairs
- Bands
- Aircraft Maintenance
- Aviation Operations
- Civil Affairs
- Electronic Maintenance and Calibration
- General Engineering
- Chemical
- Topographic Engineering
- Medical
- Mechanical Maintenance
- Electronic Warfare/Intercept Systems Technology
- Military Intelligence
- Signals Intelligence/Electronic Warfare Operations
- Military Police
- Infantry
- Combat Engineering
- Field Artillery
- Air Defense Artillery
- Special Forces
- Armor
- Air Defense System Maintenance
- Land Combat and Air Defense System Direct and General Support Maintenance
- Psychological Operations
- Visual Information
- Signal Maintenance
- Signal Operations
- Record Information Operations
- Ammunition
- Supply and Services
- Petroleum and Water
- Food Services
- Transportation

Source: Headquarters, Department of the Army, 1992.

The second concern was that several different Army occupational structures had been constructed in recent research projects. These structures differed from the two operational structures and from each other. For example, Hoffman (1987) and Rosse, Borman, Campbell, and Osborn (1983), developed a 23-job family structure as the first step in a large selection and classification study, known as Project A. In contrast, Johnson and Zeidner (1997) suggested that 66 job families provide significantly higher levels of selection and classification efficiency (in both statistical and practical terms) than smaller structures with 16 or 25 job families. ARI questioned the differences between the operational and empirically-based occupational structures, and among the empirical job families.

The final concern addressed the implications of recent political and technological changes for the design of Army jobs. The structure and processes of Army MOS have begun to change because of the redefinition of the post-Cold War mission of the U.S. military Services, the increased use of high technology equipment, and advances in telecommunications. These events have led ARI researchers to ask whether the Army's current occupational structure will be adequate in the twenty-first century, and whether industrial/organizational psychology (I/O psychology) has reliable and valid techniques for evaluating job design, and, if necessary, for restructuring jobs.

This series of questions guided the initial research and development of the CRVM, which is reported in Statman et. al. (1994) and Statman (1996).

The Problem

Statman et al. (1994) provide a review of job family research in private industry and in the military. Their overall conclusion was that the structure of occupations varies as a function of three variables:

- the purpose of the research or the operational application (e.g., employment test validation, structuring of career ladders, development of performance appraisal systems, design of training) (Pearlman, 1980);

- the type of data used to create the occupational structures (e.g., tasks, aptitude requirements, global duties and responsibilities) (Colbert & Taylor, 1978; Taylor, 1978; Taylor & Colbert, 1978; Ballentine, Cunningham, & Wimpee, 1992; and Reynolds, Laabs, & Harris, 1996; and

- the technique for grouping data (including rational methods like the Q-Sort technique and empirical procedures like Ward's hierarchical cluster analysis, k-means partitioning, and others) (Zimmerman, Jacobs, & Farr, 1982; Harvey, 1986; Garwood, Anderson, & Greengart, 1991).

Additionally, the researchers found that there are no generally accepted methods for validating occupational structures. Most published occupational classification projects included some type of external validation of job structure, but often validity was based on researcher judgement. Few studies examined reliability (Zimmerman, et al., 1982). Statman et al. (1994) concluded that the absence of empirical validation studies for occupational structures is largely due to the unavailability of statistical techniques for clustering procedures. They proposed a three-stage model, the CRVM, for measuring the accuracy of cluster structures and recommended that statistical hypothesis testing techniques be investigated. Statistical tests of cluster accuracy are rarely part of clustering techniques and validation methods because little is known about the population distributions of the indexes in data appropriate for clustering. Although the CRVM was developed as a technique for I/O psychologists and occupational analysts, it is a general approach to the problem of clustering data and can be used in other social and physical sciences with a broad range of data types. The CRVM consists of three procedures that measure:

- internal validity,
- reliability (or consistency), and
- congruence of cluster structures.

The design of the CRVM was based on a review of the cluster analysis and numerical taxonomy literatures. It is a compilation of procedures and indexes that have been developed and tested with synthetic data bases (Hubert & Arabie, 1985; McIntyre & Blashfield, 1980; Milligan, 1981a; Milligan & Cooper, 1985; Milligan & Schilling, 1985; Milligan, Soon, & Sokol, 1983). We modified some of the procedures in this project and added a statistical technique for testing hypotheses. Our general approach was suggested by Jain and Dubes (1988) and others (Milligan, 1980; Milligan, 1981b; Milligan & Cooper, 1987), but to our knowledge it has not been developed and evaluated with empirical data until now.

Overview of the Cluster Reliability and Validity Method (CRVM)

Cluster analysis is a data reduction technique and a research tool for understanding the latent structure of empirical data in many basic and applied science disciplines (e.g., biology, marketing, geophysics, medicine, meteorology, anthropology, geography, and psychology). At present there are no comprehensive, statistically-based procedures for evaluating the reliability and validity of cluster structures. Thus, cluster analysis is limited to being, for the most part, an exploratory data analysis technique. The CRVM was developed to fill this gap.[2] It integrates separate strands of cluster validation research by including measurement of the three basic concepts of cluster structure accuracy, which we define below.

Internal validity. The measurement of the internal validity of the cluster structure of a set of observations or objects involves identifying the number and composition of clusters that provides the best representation of the underlying relationships among the objects (Jain & Dubes,

[2]There is only one widely available commercial clustering statistic, SAS's Cubic Clustering Criterion (CCC). However, that procedure is limited to internal validation and can be used with only a fairly narrow range of data types (i.e., orthogonal variables). Further, the CCC is quite difficult to evaluate and has no intuitive meaning or interpretation. In comparisons of CCC against Hubert's gamma, the internal validity index used in this project, Gamma was much easier to interpret (Statman et al., 1994). An independent study by Milligan (1981) found that Gamma was more accurate than CCC.

1988; Milligan, 1981a). This has been called the fundamental problem in cluster analysis, and should be the first step in evaluating the quality of a cluster solution obtained by empirical or rational means.

Consistency and reliability. The *consistency or reliability* of a cluster structure is its stability across alternative samples or clustering methods. Consistency analysis is a replication method for evaluating whether the observed cluster structure is a good representation of the population cluster structure. The limitation of replication studies is that a negative finding provides little or no information about the sources of inconsistency (Milligan & Cooper, 1987).

Congruence of cluster structures. Measurement of the congruence of cluster structures involves evaluating the overlap of structures obtained from different sources of data. This type of external comparison in occupational research is often limited to evaluating the overlap of new clusters with existing operational job groupings, or with other cluster structures based on different job descriptors. For example, it may be important in developing a training program for a new piece of equipment to compare the groupings of tasks according to both the abilities required to perform them successfully and how difficult or important the tasks are. Although an external comparison or measurement of congruence is not a validation, it provides diagnostic information about the extent of change that could be expected by substituting new clusters for preexisting ones, or about the similarities and differences in the definitions of job structures based on different dimensions of work (e.g., tasks, behaviors and aptitudes).

A fourth component of cluster validation is the *evaluation of a cluster structure against an external criterion.* This process was excluded from the CRVM because it must be specifically tailored in each situation to address the purpose for which the classification is being used. External validation is an important step in applied research when the clusters will be used in decision-making. The external criterion should be the effectiveness of the cluster structure for accomplishing some operational purpose, e.g., grouping similar jobs together for development of a performance appraisal system (Sackett, 1988).

Overview of the Cross-Validation and Internal Validity (CV*IV) Procedure

We conducted initial tests of the CRVM procedures and indexes in Statman et al. (1994) and Statman (1996). The results showed that the techniques were feasible for use in basic and applied research and that the outputs were easy to interpret. In this study we developed a single CV*IV procedure that measures both consistency across samples (i.e., cross-validation consistency) and internal validity, and includes a statistical procedure for testing hypotheses. The general form of the hypothesis test in the CV*IV procedure is the following:

Ho: the structure of the population is random (i.e., there are no clusters); and

Ha: there are clusters in the population.

The CV*IV procedure can be used for examining a range of cluster structures to select the one with the best fit to the data. In addition, it may also be used to informally examine the significance of a single cluster structure (derived on some rational basis by the user). However, the latter application does not exactly fit the test problem described above as will be explained in the next section.

We designed the CV*IV procedure to test the null hypothesis of no clusters with a Monte Carlo procedure that generates 100 or more random samples (with either multivariate normal or uniform distributions). Our approach is to cluster the empirical sample and each of the 100 or more synthetic random samples separately. We then calculate a cluster reliability and validity index for each sample. We plot the distribution of indexes for the random samples and overlay the empirical index on this plot. Using a probability value of .05 for making a Type I error, we reject the null hypothesis of no clusters in the population, in favor of the alternative hypothesis of clusters, if the empirical value appears in the top five percent of the synthetic sampling distribution. In other words, if the value of the empirical cluster index is in the top five percent of the sampling distribution, we conclude that there is less than or equal to a 5 in 100 chance of finding a significant cluster structure in the sample when the population contains no clusters. We assume that the number and content of clusters found in the sample is the best estimate of the population cluster structure.

The remainder of this report describes the CV*IV procedure, and presents our empirical validation of it. Appendix E describes some uses of the CV*IV procedure for different types of clustering problems.

METHOD

Description of the CV*IV procedure

The CV*IV procedure is a method for using sample data to investigate whether the population contains clusters, and, if so, how many and what their content is. The number and content of the clusters obtained in the sample is considered to be the best estimate of the population cluster structure. The CV*IV procedure produces a cross-validated estimate of internal validity. If the CV*IV index is not significant, then we conclude that the population does not contain stable, internally valid clusters. If the index is significant, then we conclude that the cluster structure is internally valid and stable.

The CV*IV procedure was designed to select the cluster structure that provides the best fit with the data[3] from among a range of alternatives with 2 to n-1 clusters, where n is the number of objects or observations being clustered in cross-samples. Note that as the number of clusters in a structure changes, the content of the clusters also will change. Both the number and content of the clusters in alternative solutions impact the CV*IV results. The hypothesis test for this application of the CV*IV is:

H_o: the population is randomly distributed (i.e., there is no cluster structure in the population);

H_a: the population contains between 2 and n-1 clusters.

Another potential way to use the CV*IV procedure is to test whether a specific number of clusters and configuration of objects within clusters reflects the population structure. This hypothesis test should not be employed unless the user has a well-justified rationale for choosing a specific number of clusters. We cannot interpret results based on this analysis in the same fashion as in the hypotheses testing problem given above. We tentatively suggest the following hypotheses:

H_o: the population is randomly distributed;

H_a: the population consists of k clusters.

We do not recommend using the CV*IV procedure for the second type of analysis except under special circumstances. We mention it here because we suspect that users will be tempted to form this type of research question and we want to raise the problems from the start.

[3] The data are represented by the proximity matrix used in the clustering algorithm.

The difficulty with the second approach to identifying cluster structure is that several structures can be statistically significant--but not optimal. In other words, if there is a non-random cluster structure in the data, then a range of numbers and configurations of clusters will probably be statistically significant, but only one structure will provide the optimal fit. For example, say the population has 6 clusters. Chances are that 4-, 5-, and 7-cluster structures will provide better than random fits with the data. This will result in significant CV*IV index values for 4- to 7-cluster solutions. However, the value of the CV*IV index will be highest for the 6-cluster structure.

If the user does not examine a range of cluster structures, which differ in the number and configuration of clusters, then he or she might select a significant, but non-optimal structure. This might be all right under certain circumstances, e.g., when the constraints of the situation for which clusters are being formed requires a certain number of reliable, valid (although not necessarily optimal) clusters.

The most serious problem arises if the user's educated guess about the number of clusters is way off the mark as in the 9-cluster example presented above. Again, we do not recommend the second hypothesis testing procedure, except when the user has a sound rational basis for believing that a single cluster structure may fit the data, because more than one cluster structure can be statistically significant--but not optimal.

The CV*IV Index: Hubert's Gamma

Milligan (1981a) and Milligan and Cooper (1985) examined the properties of 30 indexes for measuring the internal validity of cluster structures and selecting the optimal number of clusters. Hubert's Gamma was among the best in a range of conditions. We selected this index for the CV*IV procedure for this reason and because it is easy to interpret.

Gamma is an intuitively pleasing measure of internal cluster structure validity because in standardized form it is the sample correlation between the cluster structure matrix and the proximity matrix for a set of objects (Jain & Dubes, 1988). In other words, Gamma measures the goodness of fit between the groupings of objects in the cluster solution and the numerical estimates of proximity or distance (squared Euclidean distance in this procedure) between all possible pairs of objects.

The numerator of Gamma is the difference between consistent cluster memberships and inconsistent memberships for all pairs of objects. A consistent pair of objects occurs when objects that are assigned to the same cluster have smaller distances than objects assigned to different clusters. Inconsistency occurs when objects in the same cluster have larger distances than objects in different clusters. The denominator is the total number of unique object pairs, or n(n-1)/2, where n is the number of objects.

Gamma ranges in value from -1 to +1, and is corrected for chance matches in the two matrices. Values of Gamma are quite easy to interpret. Gamma is 1.0 when a cluster solution is perfectly consistent with the underlying data matrix, and 0.0 when pairs match by chance. In other words, a value of 0.00, or fairly close, means there is no relationship between the cluster structure and the proximity matrix. A Gamma of 1.00, or close to it, indicates a strong congruence between the cluster structure and the underlying proximity matrix.

When an empirical clustering algorithm is used, the range of Gamma is 0.00 to +1.0, because these algorithms are optimization routines designed to maximize the similarity among objects in a cluster according to some mathematical definition of similarity (e.g., minimizing the within cluster variance, or the average distance among objects within each cluster). Gamma will rarely, if ever, be 0.00 because empirical clustering algorithms identify weak patterns in any data set, including random data.

Negative values of Gamma could be obtained by chance if objects were randomly assigned to clusters. A negative value could also appear if a rational grouping strategy were used where judges were instructed to maximize the heterogeneity of objects in the clusters. However, the utility of such an exercise would be highly questionable.

How the CV*IV Procedure Works

The CV*IV procedure is a modification of a cross-validation procedure developed by McIntyre and Blashfield (1980). The McIntyre and Blashfield procedure measures cross-sample stability and the accuracy of a cluster structure in representing the "true" population structure of synthetic data. The main limitation of the McIntyre and Blashfield procedure is that it does not provide a measure of accuracy for real data. We think the CV*IV procedure is an improvement over the original method because it measures both cross-sample reliability and internal validity (i.e., the goodness of fit between the cluster structure and the proximity matrix from which it was derived). Internal validity is a method of measuring accuracy for real data.

The CV*IV procedure for selecting the optimal cluster structure has eight steps. Table 3 presents a quick reference guide to the steps. A more detailed description of the process is also provided.

Table 3. Quick Reference Guide to the CV*IV Procedure.

Step 1. Randomly divide the total sample into Cross-Samples A and B.

Step 2. Cluster Sample A.

Step 3. Use Sample A centroids to cluster Sample B.

Step 4. Conduct Steps 2 and 3 for a range of structures that can vary in the number of clusters from 2 to n-1, where n is the number of objects or observations in Sample B.

Step 5. Select the cluster solution with the highest CV*IV Gamma value and conduct statistical significance test, based on one of the following hypotheses.

a. To examine a range of cluster structures, test:

H_o: the population is randomly distributed (i.e., there is no cluster structure in the population);

H_a: the population contains between 2 and n-1 clusters.

b. To examine a single cluster structure, test:

H_o: the population is randomly distributed;

H_a: the population contains k clusters.

Step 6. Evaluate the significance level of sample gamma and reject or do not reject H_o.

Step 7. If the cluster structure is significant, recombine the cross-samples; cluster the full sample, forming the number of clusters that was determined to be statistically optimal; and define the population cluster content based on the full sample.

Step 8. Conduct additional qualitative and quantitative diagnostic analyses of the cluster structure in the full sample and make adjustments to the number and content of the clusters as necessary.

Step 1: Randomly divide the total sample into cross-samples A and B. We use a 50/50 split. Other divisions are possible (e.g., 60/40, 70/30), although we have not yet investigated the effects of these alternative splits on the results of the CV*IV procedure.

Step 2: Cluster Sample A. We use the Ward hierarchical cluster analysis (HCA) procedure because it has performed favorably in numerous studies of occupational data (e.g., Alley, Treat, & Black, 1988; Garwood et. al., 1991). Further, the Ward minimum variance algorithm fits well with the concept of correlation in Hubert's Gamma, our CV*IV index. The CV*IV procedure can easily be modified to accommodate other clustering algorithms and internal validity indexes, as long as they are conceptually congruent with the data being clustered and the validation approach of the CV*IV procedure itself.

Note that it is important to select a clustering algorithm and index of internal validity that make sense in terms of the data being clustered. Each clustering algorithm represents a specific mathematical definition of a cluster (Aldenderfer & Blashfield, 1984). This definition must match the researcher's or user's notion of how the objects in the population of interest form clusters. For example, the single linkage HCA algorithm forms clusters by adding a new object to the cluster which contains the object to which the new object is most similar, i.e., from which it has the smallest distance. In other words, only a single linkage within a cluster is needed to add a new member. This algorithm tends to form chain-like clusters.

In comparison, the Ward hierarchical minimum variance technique forms clusters that minimize the within-groups sum of squares or error sum of squares. The Ward method tends to form spherical clusters of equal size. As another example, the average linkage HCA method adds an object to the cluster for which it minimizes the average distance between all pairs of objects. This algorithm produces results that are fairly similar to the Ward method, but it tends to produce a larger number of clusters--some large and some quite small in size.

The choice of proximity index (either a measure of distance like Euclidean distance, or a measure of similarity like the correlation coefficient) should also be selected to reflect the conceptual relationships among the objects. The proximity index (we used squared Euclidean distance, which is usually the default for the Ward procedure) is computed from the n x p raw data file, where n is the number of objects (observations) in the cross-sample and p is the number of variables on which the objects have been measured. The proximity measure indicates the strength of the relationship between any two objects. The proximity matrix is an n x n matrix.

Step 3: Use Sample A cluster centroids to cluster Sample B and compute the goodness-of-fit index, Hubert's gamma.

Step 4: If examining several cluster solutions, conduct Steps 2 and 3 for a range of numbers and configurations of clusters (using a single clustering algorithm). As many as n-2 cluster solutions can be examined in the CV*IV procedure. We exclude 2 cluster solutions, the single cluster structure and the structure in which each object is a cluster, because these solutions usually will not be of practical value. The user can also examine any subset of structures within the range of 2 to n-1 clusters. If the research has a hypothesis about a specific cluster structure, then he or she can skip this step, examining only the specific cluster structure.

Step 5: Select the cluster solution with the highest CV*IV Gamma value and conduct a statistical significance test. Statistical hypothesis tests are typically not available for clustering algorithms. We developed a procedure outlined by Jain and Dubes (1988), which we describe in Appendices B and C.

Step 6: Compare the observed sample gamma value to the sampling distribution of gamma for random data. When choosing the best cluster structure from a set of alternatives, test the following hypotheses:

H_o: the population is randomly distributed (i.e., there is no cluster structure in the population);

H_a: the population contains between 2 and n-1 clusters.

If sample gamma is in the top five percent of the sampling distribution:

- reject H_o at $p = .05$;
- conclude that the best estimate of the population cluster structure is the sample number of clusters.

If gamma is not in the top five percent of the sampling distribution:

- do not reject H_o at $p = .05$;
- conclude that the data (e.g., tasks, abilities, jobs, etc.) are randomly distributed in the population; i.e., there is no cluster structure among the objects.

When examining the statistical significance of a specific cluster structure, test these hypotheses:

H_o: the population is randomly distributed;

H_a: the population contains k clusters.

Note that no definitive conclusions about the population cluster structure can be drawn from this analysis.

Step 7: Recombine the cross-samples and recompute the cluster structure for the optimal number of clusters using all the data in the full sample. The user then conducts a content analysis and defines or labels the clusters using all of the information in the total sample.

Figure 1 contains sample output from the CV*IV procedure. In this example the user examines structures with between 2 and 10 clusters in a sample size of 50 (which becomes 25 in the cross-samples). The first page of output shows that the 5-cluster structure is found to be optimal and statistically significant at the p = .01 level. The second page presents the plot of the observed gamma value against the sampling distribution of gamma values (computed from 100 multivariate random normal samples).

The third output page plots the observed gamma values for all cluster solutions that were examined (in this case 2 to 10) by the number of clusters. This plot is very useful for ascertaining whether any of the non-optimal cluster structures had high values of gamma. If so, chances are some of these values may have been statistically significant, although they were not the largest. Since there are no statistical procedures at present to place a confidence interval around gamma for a given cluster solution, the CV*IV procedure includes additional output (see Step 8 below) that will help the researcher to evaluate whether the differences between two or more similar cluster structures are practically significant.

Step 8: Conduct additional quantitative and qualitative analysis of the optimal cluster structure. As mentioned above, more that one cluster structure may be statistically significant in the sample. We select the structure with the highest sample CV*IV gamma coefficient. At present there are no statistical procedures for placing a confidence interval around gamma or the number of clusters. In fact, this problem may be intractable because we cannot estimate the sampling distribution of Gamma (or other cluster reliability and validity indexes) for different population cluster structures. Therefore, we think it is important to permit the user to evaluate alternative structures similar to the optimal structure. The purpose of these analyses is to bring expert user judgement into the clustering process and to aid this judgement with additional qualitative and quantitative tools.

Our approach is to compare the optimal k-cluster structure with structures having one fewer (k-1) or one more (k+1) cluster. We use the Rand (1971) simple matching coefficient, and a Rand contingency table that compares the object-cluster memberships of two cluster structures, to perform these analyses.

Tables 4 and 5 present sample CV*IV output for diagnostic analysis of the data in Figure 1. The Rand coefficients and contingency tables for comparison of the optimal 5-cluster structure with the 4- and 6-cluster solutions, respectively, are provided.

The Rand coefficient measures the degree of overlap between two (binary) cluster structure matrices. It has values between 0 and 1. We corrected it for chance using the Hubert and Arabie (1985) correction for chance matching. If two cluster structures have the same number of clusters (which they will not in our application) and there is a perfect match between the two structures, the Rand value will be 1. If there is no congruence between the two structures, then the Rand will be 0.

Summary of Cluster Structure Evaluation

Observed Gamma Indices

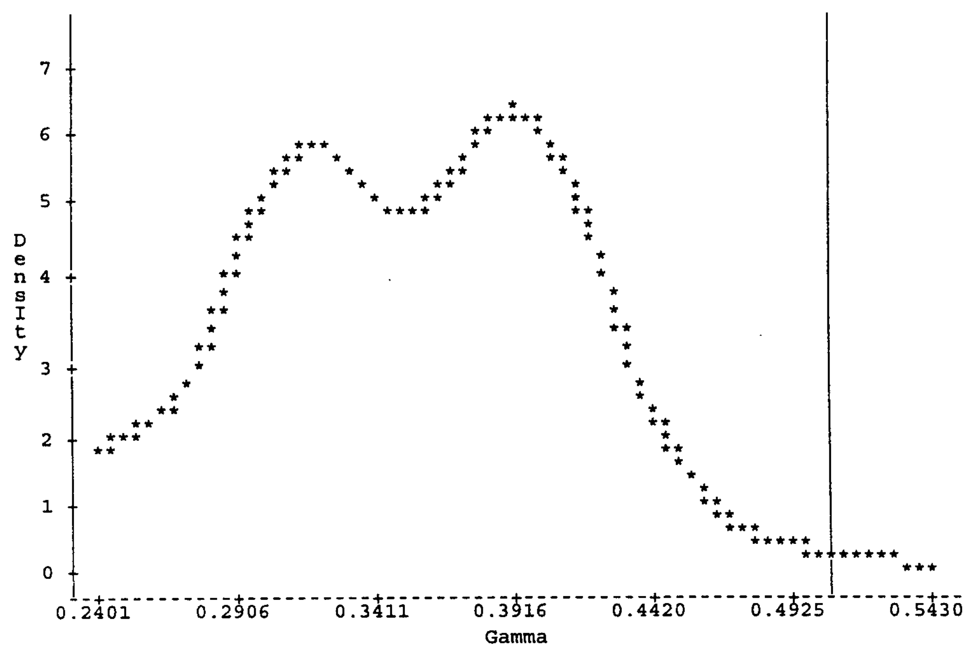
NCLUSTER	OBS_GAM
2	0.290991
3	0.437476
4	0.366399
5	0.507991
6	0.376139
7	0.363164
8	0.394496
9	0.359222
10	0.359222

Summary of the Best Cluster Solution

No. of Clusters	5
Observed Gamma	0.5079911
MC Mean of Gamma	0.352842
MC S.D. of Gamma	0.0035932
P-Value	0.01
No. of MC Samples	100

Distribution of Gamma
Under the Null Hypothesis of No Cluster Structure
Based on NORMAL Distribution

Structure = 5-Cluster Solution
Gamma Index = 0.50799 (Reference Line)
Approximate P-value = 0.0100



NOTE: 88 obs hidden.

Figure 1. Sample CV*IV Output

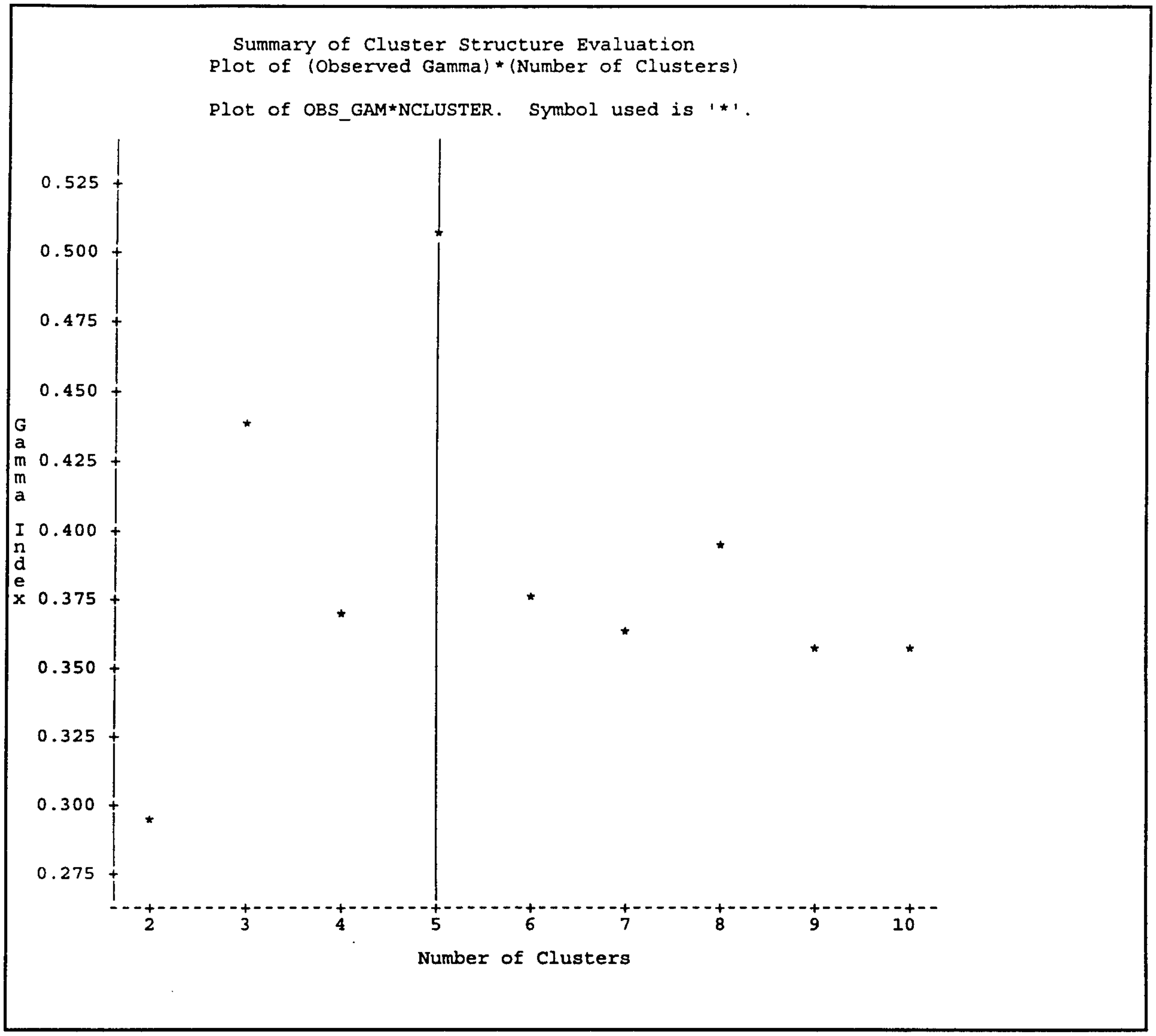


Figure 1 continued. Sample CV*IV Output

Table 4. Comparison of 4- and 5-Cluster Structures for Sample CV*IV Output

Rand Contingency Table Comparing

5-Cluster and 4-Cluster Solutions
CORRECTED RAND = 0.712794

	4-Cluster Solution				Row Totals
	CL1	CL2	CL3	CL4	
5-Cluster Solution					
CL1	15	0	0	0	15
CL2	0	10	0	0	10
CL3	0	0	8	0	8
CL4	0	0	0	8	8
CL5	9	0	0	0	9
Column Totals	24	10	8	8	50

Table 5. Comparison of 5- and 6-Cluster Structures for Sample CV*IV Output

Rand Contingency Table Comparing
5-Cluster and 6-Cluster Solutions
CORRECTED RAND = 0.960455

	6-Cluster Solution						Row Totals
	CL1	CL2	CL3	CL4	CL5	CL6	
5-Cluster Solution							
CL1	15	0	0	0	0	0	15
CL2	0	10	0	0	0	0	10
CL3	0	0	8	0	0	0	8
CL4	0	0	0	5	0	3	8
CL5	0	0	0	0	9	0	9
Column Totals	15	10	8	5	9	3	50

Examination of Tables 4 and 5 shows that the 5- and 6-cluster structures are a better match than the 4- and 5-cluster solutions. The Rand is .96 for the 5-6 comparison and .71 for the 4-5 comparison. Table 4 shows that Cluster 1 (the largest cluster) in the 4-cluster structure is split into two fairly equal clusters in the 5-cluster structure. Table 5 shows that Cluster 4 (note that order of clusters does not matter in computing Rand) in the 5-cluster structure (a relatively small cluster) is divided into two fairly small clusters in the 6-cluster solution, making these two structures highly similar.

After examining the table of object-cluster memberships for the 4-, 5-, and 6-cluster solutions (not shown here), the user would probably reject the 4-cluster structure. However, the user would have to make an expert judgement, based on knowledge of the data, about whether the 5- or 6- cluster structure was more useful for his or her purpose.

The final section of this report presents additional examples covering the two applications of the CV*IV procedure. In addition to the Rand analysis, the CV*IV procedure provides the following descriptive information on the obtained cluster structure in the full sample:

- object-cluster memberships,
- distance of each object from the cluster centroid,
- cluster distance statistics,
- cluster centroids and standard deviations, and
- cluster sizes.

In summary, the CV*IV procedure is a statistical approach for measuring cluster structure internal validity and cross-sample stability. It tests a sample cluster structure against a sampling distribution from a synthetic random population. The CV*IV procedure answers the question about whether the population contains clusters or not. The number of clusters and the cluster composition found in the sample are taken to be the best estimates of the population cluster structure based on a specific clustering algorithm and set of input data. It is important to repeat here the findings summarized in Chapter One, that cluster structures are dependent upon the purpose for the clustering, the type of data, and the definition of a cluster inherent in a clustering algorithm.

The CV*IV procedure was designed to be a statistical tool to aid the user in making decisions about cluster structure. However, it should not be the only piece of information evaluated. The diagnostic analyses provided in the CV*IV procedure are designed to help the researcher determine the utility of the obtained cluster structure. Other analyses should also be conducted, e.g., a congruence analysis that compares the obtained cluster solution to other meaningful structures, if available; and an external validation study. Ideally, an external validity study should always be conducted to obtain quantitative and/or qualitative measures of utility for a specific purpose.

EXPERIMENTAL VALIDATION OF THE CV*IV PROCEDURE

Overview of the Experimental Design

We conducted two separate tests of the CV*IV procedure. The objective of the first evaluation was to examine whether our Monte Carlo-based statistical test for the Type I Error actually produced a five percent error rate, as we designed it to do. The objective of the second evaluation was to examine the power and precision of the CV*IV procedure in samples of real data having an a priori defined cluster structure.

Validation of the Type I Error Rate in CV*IV Statistical Procedure

The purpose of this set of analyses was to evaluate how well actual Type I error rate is preserved by our Monte Carlo procedure. The Type I error is defined as rejecting the null hypothesis when it is true. For the CV*IV procedure, a Type I error would be to conclude that the population contains clusters when it does not.

Method for Evaluating the Type I Error Rate

To evaluate the Type I error we created a synthetic population data base that was known to be random (i.e., without clusters). We could not conduct this analysis with our test bed data set, which was the finite population of Army jobs (MOS), because, as we describe below, we assumed the Army population contained a 6-cluster structure.

We used the Monte Carlo simulation technique described in Appendix B to create two random population data bases. One random population was multivariate normal; the other was multivariate uniform. Our Monte Carlo procedure produced synthetic random data with either a normal or uniform distribution and the same statistical properties (i.e., means and variance-covariance matrix for the normal population, and means and ranges for the uniform population) as the real data. Our synthetic random populations had the same number of objects (N) and variables (p) as the Army finite population data base.

Our approach was to repeatedly sample from the synthetic random populations and to use those samples of random data as the observed samples in the CV*IV procedure. We conducted the experiment separately for the multivariate normal and uniform random populations. In both cases we selected 50 samples of size 50 from the population. Since the sample sizes were small, $N = 50$, the test of the actual p-value of the CV*IV procedure was more stringent than if moderate or large samples had been used. For each repeated application of the CV*IV procedure, the null and alternative hypotheses tested were the same as usual, i.e.,

H_o: the population is random (no clusters), and

H_a: the population has clusters.

To evaluate the actual Type I error rate in this validation procedure we repeatedly sampled from the same finite random population. Therefore, each time we rejected the null hypothesis (p-value < .05), we were actually committing a Type I error. An estimate of the actual Type I error rate was provided by the proportion of samples that led to rejection of the null hypothesis. If the CV*IV was performing at the desired significance level, say .05, we expected this proportion to be close to .05.

To formally evaluate if the observed Type I error rate of the CV*IV procedure with significance level .05 may be reasonably obtained, we use a two-sided Z-test with null and alternative hypotheses as follows:

$$Ho:\alpha=.05 \text{ vs } Ha:\alpha\neq.05$$

The test statistic in this case would be:

$$Z = (p-.05)/\sqrt{(.05x.95/50)}$$

where p is the observed Type I error rate in repeated samples.

In essence, what we have done in this validation of the Type I error is to conduct a test of proportion on the Type I error rate observed from the CV*IV hypothesis testing procedure in the repeated samples. Not rejecting the null hypothesis of $Ho:\alpha=.05$ indicates that the CV*IV procedure is performing reasonably using the Type I error rate at level .05.

Results of Evaluation of the Type I Error Rate

Table 6 presents the results of the test of proportion analysis described above. We observed that in 2% of the repeated samples (1 out of 50), the CV*IV procedure incorrectly rejected the null hypothesis of randomness at .05 significance level. We computed the Z-value of p = .02 using the formula shown above and obtained -0.97. The p-value for Z = -0.97 for a two-sixed test is 0.3320. Therefore, we cannot reject the null hypothesis that the true Type I error rate of the CV*IV procedure is .05.

Table 6. Hypothesis Test Results for Type 1 Error

Observed Proportion p	0.02
Z-value	-0.97
p-value	0.3320
Conclusion	Do Not Reject Null

We only present the analysis for the multivariate random normal population in Table 6. The results for the uniform population were comparable. Thus, we concluded that our CV*IV Monte Carlo statistical test preserves the desired Type I error rate of .05.

Analysis of the Power and Precision of the CV*IV Procedure

The purpose of these analyses was to evaluate the accuracy of the CV*IV procedure in real data having a known cluster structure. We developed the CV*IV procedure to overcome two weaknesses of current cluster analysis technology.

Our first objective was to implement a method for measuring cluster structure reliability (consistency) and internal validity in real data. Most cluster reliability and validity techniques have been developed and tested in synthetic data. Therefore, their utility for analysis of real data is limited at best. Our second objective was to develop a statistical test for the reliability and validity procedure so that cluster analysis techniques could be moved from the realm of exploratory data analysis to the domain of inferential statistics.

In the validation of the statistical test described above, we examined the accuracy of the CV*IV procedure concerning Type I errors. Now we turn to the accuracy of the CV*IV in detecting cluster structure when it is there (i.e., in the population). Since we designed the CV*IV procedure for use with real data, we decided that it must be validated in real data. However, this presents two problems. It is rarely possible to obtain real population data and it is usually impossible to know the structure of the population. We were able to address these two problems because we had access to the finite population of Army jobs, which is described below.

The Army population data base was compiled in the Joint Service Job Performance Measurement Project (Harris, McCloy, Dempsey, Roth, Sackett, & Hedges, 1991), which included entry-level military jobs across all four Services. Harris, McCloy, Dempsey, DiFazio and Hogan (1993) used the Army jobs in a subsequent study of alternative selection and classification models. Only the Army jobs were examined in the present study.

Job descriptors were obtained for 263 Army MOS using job analysis information from the Dictionary of Occupational Titles (DOT) (Harris et al., 1991). The DOT data base of occupational codes and job analysis ratings on 44 items was obtained for civilian jobs from the National Technical Information Service (U.S. Department of Labor, 1977). These jobs were matched to all entry-level Army jobs in existence in the mid-1980's using a military-civilian crosscode data base (Lancaster, 1984; Wright, 1984). Several MOS, e.g., many electronics jobs, received identical descriptors because the civilian job structure was not as differentiated as the Army MOS.

The 44 DOT items cover worker functions (the DOT data, people, things scales), training time, cognitive aptitudes, temperaments, interests, physical demands and working conditions. We reduced the items to four Army-specific orthogonal principal components, rotated to varimax simple structure. The principal components accounted for about 50 percent of the variance in the job descriptors and were labeled: 1) working with things, 2) complexity, 3) work environment, and 4) dealing with people and stressful working conditions. This DOT data base was considered to be the population of Army jobs at data collection in the late 1980's.

Method for Determining the Power and Precision of the CV*IV Procedure

Determining the cluster structure of the population in real data. Our method for determining the structure of the Army population of 263 jobs was to apply the Ward HCA technique, which we use in the CV*IV procedure, to construct cluster structures containing from 2 to 262 clusters, based on the DOT data. We computed the Gamma coefficient for each structure and selected the structure with the highest Gamma.

Figure 2 shows the plot of Gamma by number of clusters. Appendix D, Table 1, shows the values of Gamma for structures with 2 to 262 clusters. The 6-cluster solution provided the best fit with the population proximity matrix. However, the Gamma for the 5-cluster structure was very similar. To supplement our internal validity analysis we also examined the Rand values for comparison of the 5- and 6-cluster structures to the 4-, 7- and 8-cluster structures (see Appendix D, Table 2), and the Rand contingency table for the 5-6 cluster comparison (see Appendix D, Table 3). We also compared the 5- and 6-cluster structures in terms of cluster content, centroids, and distances (see Appendix D, Tables 4 and 5).

Based on these analyses, we decided that the 5- and 6-cluster structures were equivalent and that they best described the population cluster structure. The main difference between the two solutions was that the larger structure split combat jobs into a separate cluster. These jobs differed from the larger category of unskilled jobs only on the third factor--inside or outside working conditions.

We expected that if valid, the CV*IV procedure would tend to identify 5- or 6- cluster structures as optimal in samples under a range of conditions. We tested this hypothesis in the analyses which follow.

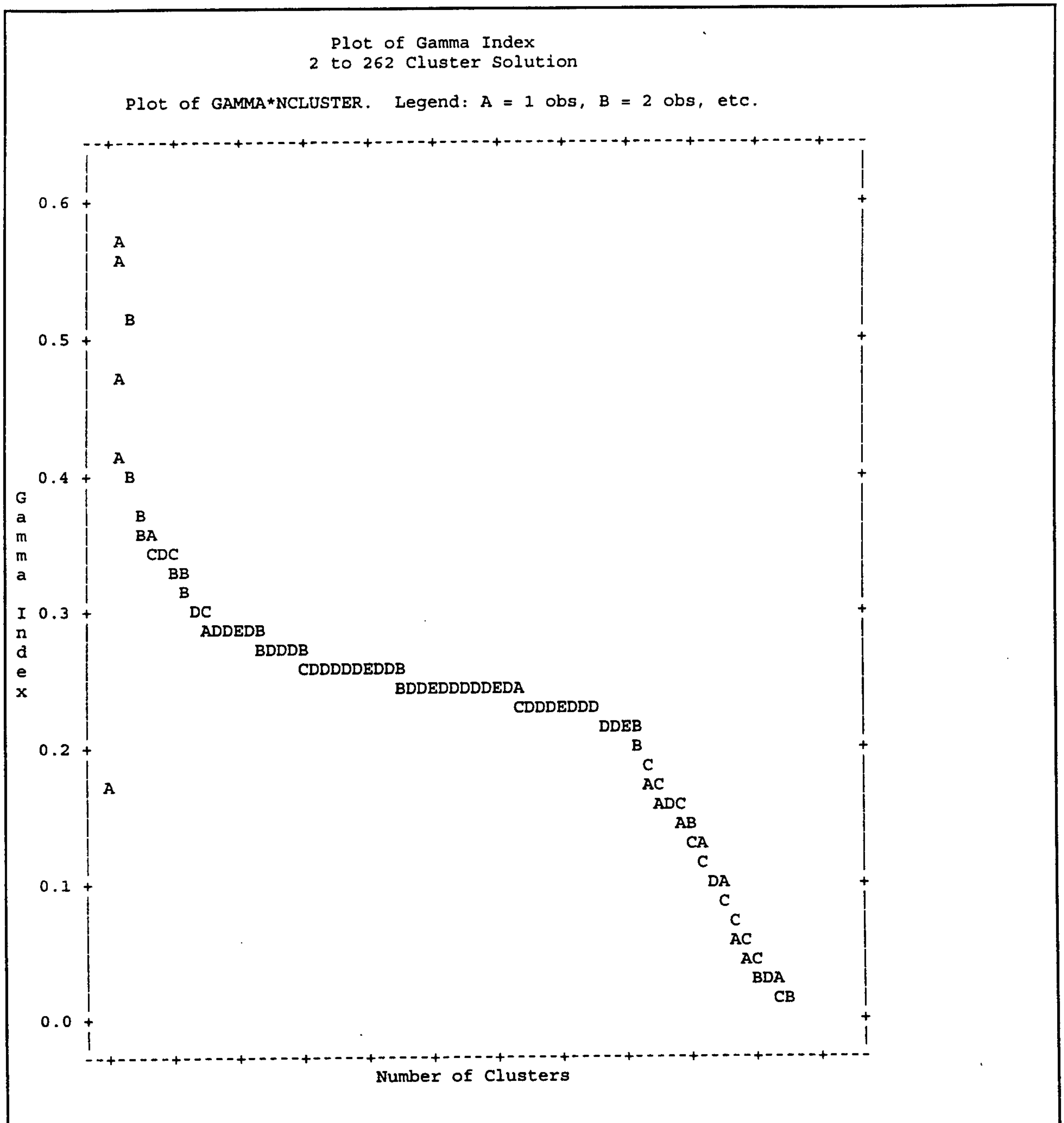


Figure 2. Plot of Gamma Values by Number of Clusters in the Population

Evaluation of the Power and Precision of the CV*IV Procedure. We defined the power of the CV*IV statistical test as the percentage of sample results having a significant cluster structure (irrespective of the number of clusters selected as optimal) in repeated sampling from the Army population.

We defined precision as the proportion of sample results having a 5- or 6-cluster structure and varied the experimental conditions along two dimensions: sample size and null hypothesis. Sample size ranged from 50 to 100 and 150 before division into cross-validation samples. Two null hypotheses were investigated: the multivariate random normally distributed population and the multivariate random uniformly distributed population. Our hypotheses were that the CV*IV procedure would be more precise with large as opposed to small samples, and that the normal hypothesis would be easier to reject than the uniform null hypothesis.

These hypotheses were tested with a loglinear analysis, which included a third variable, number of clusters identified as optimal in a particular sample by the CV*IV procedure. We conducted 50 replications in each of the six experimental cells, for a total of 300 replications. Each replication entailed selecting a random sample from the finite population of Army jobs and using the CV*IV procedure to select the best cluster structure.

Results of Analysis of Power and Precision

Table 7 contains the results for the 300 samples (experimental observations). Between 0% and 14% of the sample CV*IV analyses led to "acceptance" of the null hypothesis of a random population without clusters. The inability to reject the null hypothesis of no cluster structure, when the population is known to contain clusters, is a Type II error. The inverse of the probability of making a Type II error is referred to as the power of the statistical test. The power of the test indicates the ability to detect nonrandom effects when they exist in the population. We found that the power of the CV*IV procedure ranged from 86% for sample size 50 to 100% for sample size 150 for both normal and uniform null populations.

The range of cluster structures selected as optimal by the CV*IV procedure across all samples was 2 to 10, with 63% of the observations producing structures with 5 or 6 clusters. The 5-cluster structure was selected more often than the 6-cluster structure, although the 6-cluster structure was optimal in the population. As mentioned above, the Gamma values for the 5- and 6-cluster structures in the population were very close. The Rand matching coefficient between the two structures was .95 in the population.

Examination of the Rand contingency table in Appendix D shows that the additional cluster (Cluster 2) obtained in the larger structure had relatively few jobs (17) compared to the number of jobs (41) in the most closely related cluster (Cluster 1). Upon sampling, very few of the Cluster 2 jobs would be selected. Consequently, their effect in the clustering procedure would be fairly weak. Most of the time these few jobs would be grouped in with the larger set of unskilled labor jobs, thus, resulting in a 5-cluster solution in samples.

Table 7. Results for Empirical Power Analysis of CV*IV Procedure

Summary of Optimal N-Cluster Solutions
Using 50 Replicates of Sample Size 50
Normal and Uniform Null Distribution

No. of Clusters	Normal Distribution		Uniform Distribution	
	Total	%	Total	%
1	7	14.00	7	14.00
2	0	0	2	4.00
3	1	2.00	3	6.00
4	5	10.00	4	8.00
5	18	36.00	13	26.00
6	7	14.00	9	18.00
7	7	14.00	7	14.00
8	3	6.00	2	4.00
10	2	4.00	3	6.00

Summary of Optimal N-Cluster Solutions
Using 50 Replicates of Sample Size 100
Normal and Uniform Null Distribution

No. of Clusters	Normal Distribution		Uniform Distribution	
	Total	%	Total	%
1	0	0	1	2.00
3	0	0	2	4.00
4	3	6.00	9	18.00
5	27	54.00	19	38.00
6	7	14.00	9	18.00
7	10	20.00	4	8.00
8	3	6.00	4	8.00
9	0	0	1	2.00
10	0	0	1	2.00

Table 7 continued. Results for Empirical Power Analysis of CV*IV Procedure

Summary of Optimal N-Cluster Solutions
Using 50 Replicates of Sample Size 150
Normal and Uniform Null Distribution

	Normal Distribution		Uniform Distribution	
	Total	%	Total	%
No. of Clusters				
4	3	6.00	5	10.00
5	21	42.00	21	42.00
6	18	36.00	19	38.00
7	6	12.00	3	6.00
8	2	4.00	1	2.00
9	0	0	1	2.00

Although the sample results across 300 replications of the CV*IV procedure show that it is highly accurate, the findings also demonstrate the need for user judgement in deciding upon the final number and configuration of clusters for a given research or applied purpose. Since we are working with real data that contain complex relationships among the objects being clustered, including overlapping clusters as we found for the Army population, there may not be one best cluster structure. Further, since statistical procedures for setting a confidence interval around the number of clusters do not exist at this time, we must use expert judgement to evaluate the optimal cluster structure for a given purpose.

Table 8 presents the means, medians, modes, and standard deviations of the sample results by sample size and statement of the null hypothesis. The 5- and 6-cluster structures were selected most often across all conditions with very little variance. Table 9 shows the results of the loglinear analysis. Sample size and null hypothesis did not significantly impact selection of the optimal cluster structure. However, there was a significant interaction of sample size and number of clusters (k). As sample size increased, the range of k decreased from 2 -10 for small samples to 4 - 9 for large samples. Further, the percentage of samples with optimal 5- and 6-cluster structures increased from 50% for small samples to 78% for large samples. Table 10 shows the percentage of samples for which 5- or 6-cluster structures were obtained by sample size, and the range of cluster structures obtained by sample size. Note that Table 10 also shows these results for the full population (which was divided in half upon cross-validation), but that they were not included in the loglinear analysis.

Table 8. Average Number of Clusters in Optimal Solution by Sample Size and Null Distribution

Summary Statistics for Number of Clusters
Corresponding to Optimal Cluster Structure Solution
Total of 300 Replications

Distribution	Sample Size	N	MODE	MEDIAN	MEAN	STD
Normal	50	50	5	5	5.88	1.78
	100	50	5	5	5.66	1.06
	150	50	5	6	5.66	0.92
Uniform	50	50	5	6	5.74	1.94
	100	50	5	5	5.54	1.49
	150	50	5	5	5.54	0.97

Table 9. Log-Linear Analysis

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
NCLUSTER	5	96.00	0.0000
SAMPSIZE	2	0.54	0.7651
NCLUSTER*SAMPSIZE	10	21.88	0.0158
LIKELIHOOD RATIO	17	11.99	0.8005

Table 10: Percentage of samples by sample size and range of clusters

Correct *k* (5/6) selected

- 50% of the time w/N = 50
- 68% of the time w/N = 100
- 78% of the time w/N = 150
- 94% of the time w/N = 263

Range of *k* decreased as *N* increased

- $N = 50$ ($k = 2$ - 10)
- $N = 100$ ($k = 3$ - 10)
- $N = 150$ ($k = 4$ - 9)
- $N = 263$ ($k = 2$ - 7)

DISCUSSION AND CONCLUSIONS

We derived five major conclusions from the validation of the CV*IV procedure. First, analysis of the Type I error rate for the CV*IV statistical procedure demonstrated that the statistical test has high fidelity. Second, the power of the CV*IV procedure was also quite high (between 86% and 100%). Third, as expected in sample-based procedures, the precision of the CV*IV technique varied with sample size. However, it was still quite accurate in very small samples, with 50% of the replications producing 5- or 6-cluster structures. This rose to 78% with large samples. Fourth, we concluded that the model of randomness, whether multivariate normal or uniform, did not affect the clustering results.

Finally, we found that the CV*IV procedure provides useful diagnostic information for comparisons of the optimal cluster structure with alternative structures. This allows the user to incorporate expert judgement into the process of selecting the best possible cluster structure.

Limitations of the CV*IV Procedure

The CV*IV procedure has two major limitations. The first is that measurement of internal validity and cross-validity are confounded in the technique. Hubert's Gamma (corrected for chance) is used as a measure of internal validity in the CV*IV procedure since it is the correlation between the Sample B distance matrix and the Sample B cluster matrix. However, the cross-validation procedure, in which Sample A is clustered and the centroids are used to begin the clustering process for Sample B, introduces sample variance into the value of Gamma. If Hubert's Gamma were computed in the full sample (without cross-validation), it would be a simple function of internal validity and would be larger in magnitude than Gamma based on the CV*IV procedure, which reflects both internal validity and cross-sample differences. We developed the combined CV*IV procedure for a practical reason--for a practical reason, to provide the user with a single-stage method of developing reliable and internally valid cluster structures.

The second limitation is more serious and applies to the state-of-the-art of clustering procedures in general. Since we do not know the sampling distributions of cluster indexes in populations with cluster structures of given sizes and configurations, we cannot set a confidence interval around an optimal cluster structure. Consequently, we cannot use the CV*IV procedure to make fine distinctions between two cluster structures that vary only by a small number of clusters.

We attempted to partially address this problem by developing a set of quantitative and qualitative diagnostic procedures based on the Rand simple matching coefficient. The purpose of these procedures is to provide the user with decision-making tools when two or more cluster structures are very similar, as they were in this study.

Future Research

Since the study of cluster validation techniques (especially those that include statistical hypothesis testing procedures), is relatively new, there are many interesting unanswered questions and opportunities for developing new measurement techniques. We suggest a program of research that includes the creation of new cluster validation methods and the exploration of statistical questions.

Development of new cluster validation methods. Although the combined cross-validation and internal validity procedure we described in this report confounds the two estimates of cluster accuracy, this approach may capitalize on the strengths of each procedure. However, we would like to develop a technique that separates cross-sample stability and internal validity. A comparison of the confounded and separate methods would be useful. We would also like to see the development of statistical cluster congruence estimation procedures, including those that measure consistency across clustering algorithms. Other areas of research include testing the CV*IV procedure on different data bases, both within and outside of I/O psychology.

Statistical questions. Possible areas of research include: developing different procedures for generating synthetic sampling distributions (e.g., Bootstrap and Jackknife techniques and alternative Monte Carlo procedures); exploring other null distributions (e.g., the Poisson distribution); and investigating the distributional properties of Hubert's Gamma.

In conclusion, we have found the study and development of statistical cluster validation techniques to be a fascinating enterprise. Almost every time we tackle a new part of the project, many more theoretical, methodological or practical questions arise than we are able to resolve. This state of the science and technology of numerical taxonomy should keep researchers supplied with research opportunities well into the future.

References

Aldenderfer, M.S. & Blashfield, R.K. (1984). Cluster analysis. Newbury Park: Sage Publications.

Alley, W.E., Treat, B.R., & Black, D.E. (1988). Classification of Air Force jobs into aptitude clusters (AFHRL-TR-88-14). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Ballentine, R.D., Cunningham, J.W. & Wimpee, W.E. (1992). Air Force enlisted job clusters: An exploration in numerical job classification. Military Psychology, 4(2), 87-102.

Colbert, G.A., & Taylor, L.R. (1978). Empirically derived job families as a foundation for the study of validity generalization. Study III. Generalization of selection test validity. Personnel Psychology, 31, 355-364.

Garwood, M.K., Anderson, L.E., Greengart, B.J. (1991). Determining job groups: Application of hierarchical agglomerative cluster analysis in different job analysis situations. Personnel Psychology, 44(4), 743-762.

Harvey, R.J. (1986). Quantitative approaches to job classification: A review and critique. Personnel Psychology, 39, 267-289.

Headquarters, Department of the Army. (1992). Enlisted career management fields and military occupational specialties (AR 611-201 Revision). In Update 12-4: Military occupational classification structure. Washington, DC: Author.

Hoffman, R.G. (1987). Clustering Army military occupational specialties for Project A: Phase II (HumRRO Interim Report IR-PRD-87-22). Alexandria, VA: Human Resources Research Organization.

Hubert, L.J., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193-218.

Jain, A.K, & Dubes, R.C. (1988). Algorithms for clustering data. Engelwood Cliffs, NJ: Prentice Hall.

Johnson, C.D., & Zeidner, J. (February, 1997). Developing classification-efficient job families using differential assignment theory techniques. (Technical Note). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Maier, M.H., & Fuchs, E.F. (1972). Development and evaluation of a new ACB and aptitude area system (Technical Note 239). Arlington, VA: US Army Behavioral Science Research Laboratory. (NTIS No. Ad-703 134).

McIntyre, R.M., & Blashfield, R.K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 15, 225-238.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325-342.

Milligan, G.W. (1981a). A Monte-Carlo study of 30 internal criterion measures for cluster-analysis. Psychometrika, 46, 187-191.

Milligan, G.W. (1981b). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 370-407.

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159-179.

Milligan, G.W. & Cooper, M.C. (1987). Methodology review: clustering methods. Applied Psychological Measurement, 11(4), 329-354.

Milligan, G.W., & Schilling, D.A. (1985). Asymptotic and finite-sample characteristics of four external criterion measures. Multivariate Behavioral Research 20, 97-109.

Milligan, G.W., Soon, S.C., & Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 5, 40-47.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87(1), 1-28.

Reynolds, D.H., Laabs, G.J., & Harris, D.A. (1996). Occupational genealogies: Using job family characteristics in personnel research. Journal of Military Psychology, 8(3), 195-218..

Rosse, R. L., Borman, W.C., Campbell C.H., & Osborn, W.C. (October, 1983). Grouping Army occupational specialties by judged similarity. Paper presented to the Military Testing Association Convention, Gulf Shores, Alabama.

Sackett, P.R. (1988). Exploring strategies for clustering military occupations. In B.F. Green, Jr. & Alexandra K. Wigdor (Eds.), Linking military enlistment standards to job performance: report of a workshop. Washington, D.C.: National Academy Press.

Statman (Lightfoot), M.A. (April, 1996). An approach to evaluating the accuracy of job family structures. Paper presented at the 11th Annual Conference of the Society of Industrial and Organizational Psychology, San Diego, CA.

Statman (Lightfoot), M.A., Gribben, M.A., Harris, D.A., & Hoffman, R.G. (1994). Occupational Analysis and Job Structures (HumRRO Final Report FR-PRD-94-28). Alexandria, VA: Human Resources Research Organization

Taylor, L.R. (1978). Empirically derived job families as a foundation for the study of validity generalization. Study I. The construction of job families based on the component and overall dimensions of the PAQ. Personnel Psychology, 31, 325-340.

Taylor, L.R., & Colbert, G.A. (1978). Empirically derived job families as a foundation for the study of validity generalization. Study II. The construction of job families based on the company-specific PAQ dimensions. Personnel Psychology, 31, 341-353.

Zimmerman, R., Jacobs, R., & Farr, J. (1982). A comparison of the accuracy of four methods of clustering jobs. Applied Psychological Measurement, 6(3), 353-366.

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
00B	Diver	GM	General Maintenance	51	General Engineering
03C	Physical Activities Spec	ST	Skilled Technical	71	Administration
05B	Radio Operator	SC	Surveillance/Communica		
05C	Radio Teletype Operator	SC	Surveillance/Communica		
05D	EW/SIGINT Emitter ID/Locator	ST	Skilled Technical	98	Signs Int/Elect Warfare Oper
05G	Signal Security Specialist	ST	Skilled Technical	98	Signals Intel/Electronic Warfare Oper
05H	EW/SIGINT Intercept-IMC	ST	Skilled Technical	98	Signs Int/Elect Warfare Oper
05K	EW/SIGINT N-M Interceptor	CO	Combat	98	Signs Int/Elect Warfare Oper
11B	Infantryman	CO	Combat	11	Infantry
11C	Indirect Fire Infantryman	CO	Combat	11	Infantry
11H	Heavy Anti-Armor Wpns Inftryman	CO	Combat	11	Infantry
11M	Fighting Vehicle Infantryman	CO	Combat	11	Infantry
12B	Combat Engineer	CO	Combat	12	Combat Engineering
12C	Bridge Crewman	CO	Combat	12	Combat Engineering
12E	Atomic Demo Munitions Spec	CO	Combat	12	Combat Engineering
12F	Engineer Tracked Veh Crewman	CO	Combat	12	Combat Engineering
13B	Cannon Crewman	FA	Field Artillery	13	Field Artillery
13E	Cannon Fire Direction Speclst	ST	Skilled Technical	13	Field Artillery
13F	Fire Support Specialist	FA	Field Artillery	13	Field Artillery
15D	Lance Missile Crewmember	OF	Operators/Food	13	Field Artillery
15E	Pershing Missile Crewmember	OF	Operators/Food	13	Field Artillery
15J	MLRS/LANCE Ops/FireDir Spec	FA	Field Artillery	13	Field Artillery
16F	Light Air Def Artillery Crewmbr	OF	Operators/Food	14	Air Defense Artillery
16H	ADA Opertns-Intellignce Assis	OF	Operators/Food	14	Air Defense Artillery
16L	Sgt York Air Def Gun Crwmbr	OF	Operators/Food	14	Air Defense Artillery
16R	ADA Short Range Gunry Crwmbr	OF	Operators/Food	14	Air Defense Artillery
17C	Field Artlry Target Acq Spec	SC	Surveillance/Communica	13	Field Artillery
17L	Aerial Sensor Specialist	ST	Skilled Technical	96	Military Intelligence
17M	Remote Sensor Specialist	SC	Surveillance/Communica	96	Military Intelligence
19D	Cavalry Scout	CO	Combat	19	Armor
19E	M48-M60 Armor Crewman	CO	Combat	19	Armor
19K	M1 ABRAMS Armor Crewman	CO	Combat	19	Armor
21L	PERSHING Electronics Repairer	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
22L	NIKE Test Equipment Repairer	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
22N	NIKEHER MissileLauncherRepair	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
23N	NIKE Track Radar Repairer	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
23U	NIKE Radar-Simulator Repairer	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
24C	Improvd HAWK Firng Sect Mech	EL	Electronics	23	Air Defense System Maintenance
24E	Improvd HAWK Fire Contrl Mech	EL	Electronics	23	Air Defense System Maintenance
24G	Imp HAWK Inform CoorCentMech	EL	Electronics	23	Air Defense System Maintenance
24H	Improvd HAWK Fire Contrl Repr	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
24J	Improved HAWK Pulse Radar Rep	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
24K	ImpHAWK Cont-Wave Radar Repr	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
24L	ImpHAWK Launch&Mech Sys Repr	EL	Electronics	27	Land Comb/Air Def Syst Dir/Gen Supp Maint
24P	Defense Acq Radar Mechanic	EL	Electronics	23	Air Defense System Maintenance
24Q	NIKE-HERCULES Fire Contrl Mec	EL	Electronics	23	Air Defense System Maintenance
24U	HERCULES Electronic Mechanic	EL	Electronics	23	Air Defense System Maintenance
24W	Sgt York Air Def Gun Syst Mec	EL	Electronics	23	Air Defense System Maintenance
25J		EL	Electronics		
26B	Weapons Support Radar Repr	EL	Electronics	29	Signal Maintenance

(Continued)

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
26D	Ground Cntrl Approch Rdar Rep	EL	Electronics	28	Avia Comm/Electronics Systems Maintenance
26E	Aerial Surv Sensor Repairer	EL	Electronics	33	Electronic Warfare/Intercept Systems Maintenance
26F	Aerial Photo-Activ Sensor Rep	EL	Electronics	33	Electronic Warfare/Intercept Systems Maintenance
26H	Air Defense Radar Repairer	EL	Electronics	23	Air Defense System Maintenance
26L	Tactical Microwave Syst Repr	EL	Electronics	29	Signal Maintenance
26M	Aerial Surveillance Radar Repr	EL	Electronics		
26N	Aerial Surveillance Infrared Repr	EL	Electronics		
26Q	Tact Satell/Microwave Syst Op	EL	Electronics	31	Signal Operations
26R	Strategic Microwave Syst Op	EL	Electronics	31	Signal Operations
26T	Radio/TV Systems Specialist	EL	Electronics	25	Visual Information
26V	Strategic Microwave Syst Repr	EL	Electronics	29	Signal Maintenance
26Y	SATCOM Equipment Repairer	EL	Electronics	29	Signal Maintenance
27B	LandCombat SystemTestSpecial	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27E	TOW/DRAGON Repairer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27F	VULCAN Repairer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27G	CHAPARRAL/REDEYE Repairer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27H	HAWK Firing Section Reparirer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27L	LANCE System Repairer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27M	MLRS Repairer	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27N	Forwrd Area Alerting Rdar Rep	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27P	SgtYork Radar/Electron Repr	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
27Q	SgtYork Test Specialist	EL	Electronics	27	Land Comb/Air System Dir/Gen Supp Maint
31E	Field Radio Repairer	EL	Electronics	29	Signal Maintenance
31J	Teletypewriter Repairer	EL	Electronics	29	Signal Maintenance
31M	Multichannel Commo Equip Op	EL	Electronics	31	Signal Operations
31N	Tactical Circuit Controller	EL	Electronics	31	Signal Operations
31S	Field General Comsec Repairer	EL	Electronics	29	Signal Maintenance
31T	Field Systems Comsec Repairer	EL	Electronics	29	Signal Maintenance
31V	Tactical Commo Syst Op/Mech	EL	Electronics	31	Signal Operations
32D	Station Technical Controller	EL	Electronics	31	Signal Operations
32F	Fixed Ciphony Repairer	EL	Electronics	29	Signal Maintenance
32G	Fixed Crypto Equip Repairer	EL	Electronics	29	Signal Maintenance
32H	Fixed Station Radio Repairer	EL	Electronics	29	Signal Maintenance
33S	EW/Intercept Sys Repr	ST	Skilled Technical	33	Electronic Warfare/Intercept Systems Maintenance
34B	Punch Card Machine Operator	EL	Electronics	74	Record Information Operations
34C	Decen Auto Serv Supp Systm	EL	Electronics	74	Record Information Operations
34E	NCR 500 Computer Repairer	EL	Electronics	74	Record Information Operations
34F	Digt Subsc Message Switch Equip	EL	Electronics	74	Record Information Operations
34H	Auto Digt Message Switch Equip	EL	Electronics	74	Record Information Operations
34Y	Field Artlry TactFire Repair	EL	Electronics	74	Record Information Operations
35B	Electronics Instrument Repr	EL	Electronics		
35C	Automatic Test Equip Repairer	EL	Electronics	74	Record Information Operations
35E	Special Elec Devices Repairer	EL	Electronics	29	Signal Maintenance
35F	Nuclear Weapons Electronics Specialist	EL	Electronics		
35G	Biomedical Equipment Spec	EL	Electronics	91	Medical
35H	Calibration Specialist	EL	Electronics	35	Electronic Maintenance
35K	Avionic Mechanic	EL	Electronics	28	Avia Comm/Electronics Systems Maintenance
35L	Avionic Commo Equip Repairer	EL	Electronics	28	Avia Comm/Electronics Systems Maintenance
35M	Avionic Nav/FlightContEq Repr	EL	Electronics	28	Avia Comm/Electronics Systems Maintenance
35R	Avionic Special Equip Repr	EL	Electronics	28	Avia Comm/Electronics Systems Maintenance

(Continued)

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
36C	Wire System Instll/Operator	EL	Electronics	31	Signal Operations
36D	Antenna Installer Specialist	EL	Electronics		
36E	Cable Splicer	EL	Electronics		
36H	Dial/Manual Centrl Office Rep	EL	Electronics	29	Signal Maintenance
36K	Tactical Wire Operations Specialist	EL	Electronics		
36L	Trans ElectSwitchSys Rep	EL	Electronics	31	Signal Operations
36M	Wire Systems Operator	EL	Electronics	31	Signal Operations
41B	Topographic Instr Rep Spec	GM	General Maintenance	81	Topographic Engineering
41C	Fire Contrl Instru Rep Spec	GM	General Maintenance	63	Mechanical Maintenance
41E	Audio/Visual Equip Repairer	EL	Electronics	25	Visual Information
41G	Aerial Surveillance Photo Equip Repr	EL	Electronics		
41J	Office Machine Repairer	GM	General Maintenance	63	Mechanical Maintenance
42C	Orthotic Specialist	GM	General Maintenance	91	Medical
42D	Dental Laboratory Specialist	GM	General Maintenance	91	Medical
42E	Optical Laboratory Spec	GM	General Maintenance	91	Medical
43E	Parachute Rigger	GM	General Maintenance	76	Supply and Services
43M	Fabric Repair Specialist	GM	General Maintenance	76	Supply and Services
44B	Metalworker	GM	General Maintenance	63	Mechanical Maintenance
44E	Machinist	GM	General Maintenance	63	Mechanical Maintenance
45B	Small Arms Repairer	GM	General Maintenance	63	Mechanical Maintenance
45D	SP Field Artlry Turret Mech	GM	General Maintenance	63	Mechanical Maintenance
45E	M1 ABRAMS Tank Turret Mech	GM	General Maintenance	63	Mechanical Maintenance
45G	Fire Control System Repairer	EL	Electronics	63	Mechanical Maintenance
45K	Tank Turret Repairer	GM	General Maintenance	63	Mechanical Maintenance
45L	Artillery Repairer	GM	General Maintenance	63	Mechanical Maintenance
45N	M60A1/A3 Tank Turret Mech	MM	Mechanical Maintenance	63	Mechanical Maintenance
45T	BFVS Turret Mechanic	GM	General Maintenance	63	Mechanical Maintenance
46N	PERSHING ElecMechcal Repairer	EL	Electronics	63	Mechanical Maintenance
51B	Carpentry/Masonry Specialist	GM	General Maintenance	51	General Engineering
51C	Structures Specialist	GM	General Maintenance	51	General Engineering
51K	Plumber	GM	General Maintenance	51	General Engineering
51M	Firefighter	GM	General Maintenance	51	General Engineering
51N	Water Treatment Specialist	GM	General Maintenance	77	Petroleum and Water
51R	Interior Electrician	GM	General Maintenance	51	General Engineering
52C	Utilities Equipment Repairer	GM	General Maintenance	63	Mechanical Maintenance
52D	Power Generator Equip Repr	GM	General Maintenance	63	Mechanical Maintenance
52G	Transmisson & Distbution Spec	EL	Electronics	51	General Engineering
53B	Industrial Gas Prod Specialist	GM	General Maintenance	55	Ammunition
54C	Smoke Operation Specialist	GM	General Maintenance	54	Chemical
54E	NBC Specialist	ST	Skilled Technical	54	Chemical
55B	Ammunition Specialist	GM	General Maintenance	55	Ammunition
55D	Explsve Ordnance Disposl Spec	GM	General Maintenance	55	Ammunition
55G	Nuclear Weapons Maint Spec	GM	General Maintenance	55	Ammunition
55R	Ammo Stock Control&Acct Spec	ST	Skilled Technical	55	Ammunition
57E	Laundry & Bath Specialist	GM	General Maintenance	76	Supply and Services
57F	Graves Registration Spec	GM	General Maintenance	76	Supply and Services
57H	Cargo Specialist	GM	General Maintenance	88	Transportation
61B	Watercraft Operator	MM	Mechanical Maintenance	88	Transportation
61C	Watercraft Engineer	MM	Mechanical Maintenance	88	Transportation
61F	Marine Hull Repr	GM	General Maintenance		

(Continued)

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
62B	Construction Equipment Repr	MM	Mechanical Maintenance	63	Mechanical Maintenance
62E	Hvy Construction Equip Op	GM	General Maintenance	51	General Engineering
62F	Crane Operator	GM	General Maintenance	51	General Engineering
62G	Quarrying Specialist	GM	General Maintenance	51	General Engineering
62H	Concrete & Asphalt Equip Op	GM	General Maintenance	51	General Engineering
62J	General Construc Equip Op	GM	General Maintenance	51	General Engineering
63B	Light Wheel Vehicle Mechanic	MM	Mechanical Maintenance	63	Mechanical Maintenance
63D	SP Field Artilry System Mech	MM	Mechanical Maintenance	63	Mechanical Maintenance
63E	M1 Abrams Tank System Mech	MM	Mechanical Maintenance	63	Mechanical Maintenance
63G	Fuel & Elec System Repairer	MM	Mechanical Maintenance	63	Mechanical Maintenance
63H	Track Vehicle Repairer	MM	Mechanical Maintenance	63	Mechanical Maintenance
63J	Quart&Chem Equipment Repairer	MM	Mechanical Maintenance	63	Mechanical Maintenance
63N	M60A1/A3 Tank System Mechanic	MM	Mechanical Maintenance	63	Mechanical Maintenance
63S	Heavy Wheel Vehicle Mechanic	MM	Mechanical Maintenance	63	Mechanical Maintenance
63T	Bradley System Mechanic	MM	Mechanical Maintenance	63	Mechanical Maintenance
63W	Wheel Vehicle Repairer	MM	Mechanical Maintenance	63	Mechanical Maintenance
63Y	Track Vehicle Mechanic	MM	Mechanical Maintenance	63	Mechanical Maintenance
64C	Motor Transport Operator	OF	Operators/Food	88	Transportation
65B	Locomotive Repairer	MM	Mechanical Maintenance	88	Transportation
65D	Railway Car Repairer	MM	Mechanical Maintenance	88	Transportation
65E	Airbrake Repairer	MM	Mechanical Maintenance	88	Transportation
65H	Locomotive Operator	MM	Mechanical Maintenance	88	Transportation
65J	Train Crewmemeber	MM	Mechanical Maintenance	88	Transportation
67G	Utility Airplane Repairer	MM	Mechancial Maintenance	67	Aircraft Maintenance
67H	Observation Airplane Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
67N	Utility Helicopter Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
67T	Tact Transp Helicoptr Repr	MM	Mechanical Maintenance	67	Aircraft Maintenance
67U	Medium Helicopter Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
67V	Observ/Scout Helicoptr Repr	MM	Mechanical Maintenance	67	Aircraft Maintenance
67X	Heavy Lift Helicopter Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
67Y	AH-1 Attack Helicoptr Repr	MM	Mechanical Maintenance	67	Aircraft Maintenance
68B	Aircraft Powerplant Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
68D	Aircraft Powertrain Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
68F	Aircraft Electrician	MM	Mechanical Maintenance	67	Aircraft Maintenance
68G	Aircraft Structural Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
68H	Aircraft Pneudraulic Repairer	MM	Mechanical Maintenance	67	Aircraft Maintenance
68M	Aircraft Weapon Systems Repr	GM	General Maintenance	67	Aircraft Maintenance
71C	Exec Administrative Assistant	CL	Clerical	71	Administration
71D	Legal Clerk Specialist	CL	Clerical	71	Administration
71E	Court Reporter	CL	Clerical	71	Administration
71G	Patient Admin Specialist	CL	Clerical	91	Medical
71L	Administrative Specialist	CL	Clerical	71	Administration
71N	Traffic Management Coordntor	CL	Clerical	88	Transportation
71P	Flight Operations Coordinator	ST	Skilled Technical		
71Q	Journalist	ST	Skilled Technical	46	Public Affairs
71R	Broadcast Journalist	ST	Skilled Technical	46	Public Affairs
72E	Combat Telecomm Ctr Operator	SC	Surveillance/Communica	31	Signal Operations
72G	Auto Data Telecomm Ctr Oprtor	SC	Surveillance/Communica	31	Signal Operations
73C	Finance Specialist	CL	Clerical	71	Administration
73D	Accounting Specialist	ST	Skilled Technical	71	Administration

(Continued)

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
74B	Card and Tape Writer	CL	Clerical	74	Record Information Operations
74D	Computer/Machine Operator	ST	Skilled Technical	74	Record Information Operations
74F	Programmer Analyst	ST	Skilled Technical	74	Record Information Operations
75B	Personnel Admin Specialist	CL	Clerical	71	Administration
75C	Personnel Management Spec	CL	Clerical	71	Administration
75D	Personnel Records Specialist	CL	Clerical	71	Administration
75E	Personnel Action Specialist	CL	Clerical	71	Administration
75F	Personnel Info Mangmt Spec	CL	Clerical	71	Administration
76C	Equip Records & Parts Spec	CL	Clerical	76	Supply and Services
76J	Medical Supply Specialist	CL	Clerical	91	Medical
76P	Materl Centrl & Acctng Spec	CL	Clerical	76	Supply and Services
76V	Mat Storage & Handlng Spec	CL	Clerical	76	Supply and Services
76W	Petroleum Supply Specialist	CL	Clerical	77	Petroleum and Water
76X	Subsistence Supply Specialist	CL	Clerical	76	Supply and Services
76Y	Unit Supply Specialist	CL	Clerical	76	Supply and Services
81B	Technical Drafting Specialist	ST	Skilled Technical	51	General Engineering
81C	Cartographer	ST	Skilled Technical	81	Topographic Engineering
81E	Illustrator	ST	Skilled Technical	25	Visual Information
82B	Construction Surveyor	ST	Skilled Technical	51	General Engineering
82C	Field Artillery Surveyor	ST	Skilled Technical	13	Field Artillery
82D	Topographic Surveyor	ST	Skilled Technical	81	Topographic Engineering
83E	Photo & Layout Specialist	ST	Skilled Technical	81	Topographic Engineering
83F	Photolithographer	ST	Skilled Technical	81	Topographic Engineering
84B	Still Photographic Specialist	ST	Skilled Technical	25	Visual Information
84C	Motion Picture Specialist	ST	Skilled Technical	25	Visual Information
84F	Audio/TV Specialist	ST	Skilled Technical	25	Visual Information
91A	Medical Specialist	ST	Skilled Technical	91	Medical
91C	Practical Nurse	ST	Skilled Technical	91	Medical
91D	Operating Room Specialist	ST	Skilled Technical	91	Medical
91E	Dental Specialist	ST	Skilled Technical	91	Medical
91F	Psychiatric Specialist	ST	Skilled Technical	91	Medical
91G	Behavioral Science Specialist	ST	Skilled Technical	91	Medical
91H	Orthopedic Specialist	ST	Skilled Technical	91	Medical
91J	Physical Therapy Specialist	ST	Skilled Technical	91	Medical
91L	Occupational Therapy Spec	ST	Skilled Technical	91	Medical
91N	Cardiac Specialist	ST	Skilled Technical	91	Medical
91P	X-Ray Specialist	ST	Skilled Technical	91	Medical
91Q	Pharmacy Specialist	ST	Skilled Technical	91	Medical
91R	Veterinary Food Inspec Spec	ST	Skilled Technical	91	Medical
91S	Environmental Health Spec	ST	Skilled Technical	91	Medical
91T	Animal Care Specialist	ST	Skilled Technical	91	Medical
91U	Ear, Nose & Throat Specialist	ST	Skilled Technical	91	Medical
91V	Respiratory Specialist	ST	Skilled Technical	91	Medical
91Y	Eye Specialist	ST	Skilled Technical	91	Medical
92B	Medical Laboratory Specialist	ST	Skilled Technical	91	Medical
92C	Petroleum Laboratory Spec	ST	Skilled Technical	77	Petroleum and Water
92D	Chemical Laboratory Spec	ST	Skilled Technical	54	Chemical
93E	Meteorological Observer	ST	Skilled Technical	93	Aviation Operations
93F	Field Artlry Meteorlogic Spec	EL	Electronics	13	Field Artillery
93H	Air Traffic Control Tower Op	ST	Skilled Technical	93	Aviation Operations

(Continued)

APPENDIX A

MOS Titles, Aptitude Areas, and Career Management Field for
263 MOS in DOT Cluster Validation Data Base????

MOSID	MOSTITLE	AAID	AATITLE	CMFNO	CMFTITLE
93J	Air Traff Cntrl Radar Contlr	ST	Skilled Technical	93	Aviation Operations
94B	Food Service Specialist	OF	Operators/Food	94	Food Services
94F	Hospital Food Service Spec	OF	Operators/Food	91	Medical
95B	Military Police	ST	Skilled Technical	95	Military Police
95C	Correctional Specialist	ST	Skilled Technical	95	Military Police
96B	Intelligence Analyst	ST	Skilled Technical	96	Military Intelligence
96C	Interrogator	ST	Skilled Technical	96	Military Intelligence
96D	Imagery Analyst	ST	Skilled Technical	96	Military Intelligence
96H	Aerial Intell Spec	ST	Skilled Technical	96	Military Intelligence
97B	Counterintelligence Agents	ST	Skilled Technical	96	Military Intelligence
98C	EW/SIGINT Analyst	ST	Skilled Technical	98	Signals Intel/Electronic Warfare Oper
98G	EW/SIGINT Voice Interceptor	ST	Skilled Technical	98	Signals Intel/Electronic Warfare Oper
98J	EW/SIGINT NoncommoIntercept	ST	Skilled Technical	98	Signals Intel/Electronic Warfare Oper

APPENDIX B

THE DESIGN OF THE CV*IV STATISTICAL HYPOTHESIS TEST

The problem with statistical analysis of cluster structures is that little is known about the distributions of objects in the population and the concomitant distributions of cluster reliability and validity indexes. Consequently, several researchers (Jain & Dubes, 1988; Milligan, 1980, 1981a; Milligan & Cooper, 1985) recommend using Monte Carlo procedures, which create synthetic samples with given random population distributions, to provide the basis for statistical hypothesis tests. Note that at present the problem of constructing confidence intervals around cluster reliability and validity indexes, and the associated number of clusters, is intractable. This is because we know even less about the shapes of clusters in different data bases and the distributions of indexes for given numbers of clusters.

In designing the CV*IV procedure we assumed that both a multivariate normal distribution and a multivariate uniform distribution were valid models of randomness for many types of data in the social and physical sciences. Therefore, we developed separate statistical tests for each null hypothesis of randomness. The null hypotheses for the models are the following:

H_o: the population is multivariate normally distributed (i.e., there is no cluster structure in the population); or

H_o: the population is multivariate uniformly distributed (i.e., there is no cluster structure in the population).

In both cases the alternative hypothesis is:

H_a: the number of clusters in the population structure is within the set {2 to n-1 clusters}, where n is the number of objects in the cross-samples.

It may also be possible to use the CV*IV procedure to explore an alternative hypothesis about a specific cluster structure. We describe when this alternative hypothesis can be applied and why it should be used with care in the method.

The Monte Carlo random sample generation technique creates n x p matrices, where n is the number of objects in the cross-samples and p is the number of variables on which the objects are measured. Jain and Dubes (1988, pp. 158-159) provide a formula for determining the number of Monte Carlo samples that is adequate to obtain given probabilities of making a Type I error.

We selected a significance level of .05 because we felt that level would give us the proper balance between the probability of a Type I error and the power of the test. According to the above mentioned formula, 100 Monte Carlo samples should provide an actual Type I error rate that is reasonably close to the desired level of .05. We tested this with synthetic multivariate normal and uniform samples and found the Monte Carlo sampling distribution procedure to have high fidelity. The validation method and results are presented in the body of the report.

Appendix C presents the rationale for the statistical procedure and a proof that the p-value for the test is alpha.

To create the multivariate random normal distributions we use the Cholesky decomposition matrix to represent the variance-covariance structure of the empirical sample (which we assumed to be a good representation of the population). We then impose the sample variance-covariance structure on an n x p matrix of random normal deviates. We use a similar approach for creating the distributions of n x p orthogonal uniformly distributed deviates with the same means and ranges as those of the empirical sample. We generate 100 or more synthetic random samples for each model--normal and uniform. Each synthetic random sample is clustered using the CV*IV procedure, and the first 100 samples that produce the highest gamma value for the same number of clusters as obtained in the empirical sample are retained.

The statistical hypothesis test is carried out by separately comparing the observed gamma from the empirical sample to the sampling distributions of 100 gamma values created from the random normal and uniform models. We do this by computing the density functions for the sampling distributions and overlaying the observed gamma on each of the plots. If observed gamma is in the top five percent of the sampling distribution, we reject H_0 in favor of H_a. If not, we do not reject H_0 and conclude that the population from which the empirical sample was drawn is randomly distributed according to a normal or uniform model; i.e., we conclude that the population does not contain clusters.

APPENDIX C

DERIVATION OF TYPE I ERROR PROBABILITY

The cluster evaluation procedure presented in this paper may be viewed as a statistical test of

H_0: "no cluster structure" vs H_a: "with structure"

The alternative hypothesis does not specify a single N-cluster structure but instead accomodates a class of alternative cluster solutions. A closer scrutiny of the procedure as described in the preceeding sections reveals that it is composed of (n/2 - 2) component tests. We look at the maximum observed gamma index. If it is significant, then we reject the "no cluster structure" null hypothesis and our best estimate of the cluster structure is given by the corresponding N-cluster solution. But note that the maximum gamma can be any of the (n/2 -2) gammas corresponding to the 2-, 3-, ..., (n/2-1)-cluster solutions. That is, we could potentially be using any of the (n/2 -2) associated conditinal tests of significance. The actual test, corresponding to the maximum gamma, is determined only after observations are made. In the following discussion, we examine the overall level of significance of the test.

Let C_k denote the k-cluster solution, k=1,2,...,(n/2 -1). The "no structure" solution corresponds to C_1. The null and alterative hypotheses can now be stated as

H_0: $C=C_1$ vs H_a: $C \in \{C_k, k=2, 3, ..., (n/2 -1)\}$

Let X be an n×p matrix of random observations. Define $\Gamma_k = \Gamma(X;C_k)$, the gamma index corresponding to the k-cluster solution. The α-level test procedure is now given by the rule:

Reject H_0: $C=C_1$ if $\Gamma_{Kmax} \geq \Gamma_{Kmax,\alpha}$

Here, K_{max} is the index of the maximum gamma and is random and $\Gamma_{Kmax,\alpha}$ is the $(1-\alpha)\times 100$ percentile of the Γ_{Kmax} distribution. As implemented in our cluster evaluation program, the rule above is indirectly carried out by getting an estimate of the p-value of the observed maximum gamma based on a Monte Carlo simulation. The actual critical value $\Gamma_{Kmax,\alpha}$ is not calculated. The computation below of the overall level of significance proceeds by partitioning the entire cluster evaluation procedure into its component parts {reject H_0 if $\Gamma_k \geq \Gamma_{k,\alpha} \mid K_{max}=k$, k=2, 3, ..., (n/2 -1)}. The probabilities are then collected across partitions using the total probability rule.

$$
\begin{aligned}
\mathrm{P(Type\ I\ Error)} &= \mathrm{P}\left(\Gamma(\mathrm{X};\mathrm{C}_{K\max}) \geq \Gamma_{K\max,\alpha} \mid \mathrm{C}=\mathrm{C}_1\right) \\
&= \mathrm{P}\left(\bigcup_{k} \{(K_{\max}=k) \text{ and } (\Gamma(\mathrm{X};\mathrm{C}_{K\max=k}) \geq \Gamma_{K\max=k,\alpha})\} \mid \mathrm{C}=\mathrm{C}_1\right) \\
&= \sum_{k} \mathrm{P}\left(\Gamma(\mathrm{X};\mathrm{C}_{K\max}) \geq \Gamma_{K\max,\alpha} \mid K_{\max}=k, \mathrm{C}=\mathrm{C}_1\right) \mathrm{P}(K_{\max}=k \mid \mathrm{C}=\mathrm{C}_1) \\
&= \alpha \sum_{k} \mathrm{P}(K_{\max}=k \mid \mathrm{C}=\mathrm{C}_1) \\
&= \alpha
\end{aligned}
$$

That is, Type I Error rate α is preserved overall by the test. The union and summations above are taken over k=2, 3, ..., (n/2 - 1). The second summation was obtained from the first by noting that each of the component tests in the total probability expression is an α-level test. Also, there is an implicit distribution under which the probabilities above are evaluated.

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 1. Gamma Values for Population Cluster Structures

Population Gamma Index
2 to 262 Cluster Solutions

OBS	NCLUSTER	GAMMA
1	2	0.17768
2	3	0.42019
3	4	0.47390
4	5	0.55060
5	6	0.57324
6	7	0.51271
7	8	0.51259
8	9	0.40450
9	10	0.39589
10	11	0.37171
11	12	0.36516
12	13	0.35578
13	14	0.35172
14	15	0.35029
15	16	0.34917
16	17	0.34513
17	18	0.34407
18	19	0.34416
19	20	0.34410
20	21	0.34388
21	22	0.34056
22	23	0.33885
23	24	0.33690
24	25	0.33638
25	26	0.33223
26	27	0.32815
27	28	0.32822
28	29	0.32474
29	30	0.31121
30	31	0.31079
31	32	0.30165
32	33	0.29927
33	34	0.29767
34	35	0.29604
35	36	0.29476
36	37	0.29342
37	38	0.29331
38	39	0.29283
39	40	0.29226
40	41	0.29051
41	42	0.29031
42	43	0.29020
43	44	0.28874
44	45	0.28865
45	46	0.28862
46	47	0.28706
47	48	0.28669
48	49	0.28645
49	50	0.28621
50	51	0.28609
51	52	0.28518
52	53	0.28454
53	54	0.28254
54	55	0.28138
55	56	0.28092
56	57	0.27994
57	58	0.27950
58	59	0.27391
59	60	0.27287
60	61	0.27256
61	62	0.27180
62	63	0.27164
63	64	0.27029
64	65	0.27014
65	66	0.26998
66	67	0.26891
67	68	0.26841

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 1 continued. Gamma Values for Population Cluster Structures

Population Gamma Index
2 to 262 Cluster Solutions

OBS	NCLUSTER	GAMMA
68	69	0.26790
69	70	0.26738
70	71	0.26641
71	72	0.26589
72	73	0.26509
73	74	0.26433
74	75	0.26416
75	76	0.26399
76	77	0.26391
77	78	0.26374
78	79	0.26357
79	80	0.26339
80	81	0.26322
81	82	0.26304
82	83	0.26295
83	84	0.26259
84	85	0.26063
85	86	0.26055
86	87	0.26008
87	88	0.25961
88	89	0.25952
89	90	0.25943
90	91	0.25934
91	92	0.25916
92	93	0.25897
93	94	0.25869
94	95	0.25859
95	96	0.25840
96	97	0.25831
97	98	0.25770
98	99	0.25721
99	100	0.25712
100	101	0.25693
101	102	0.25663
102	103	0.25612
103	104	0.25394
104	105	0.25374
105	106	0.25365
106	107	0.25355
107	108	0.25345
108	109	0.25325
109	110	0.25315
110	111	0.25231
111	112	0.25125
112	113	0.24683
113	114	0.24619
114	115	0.24609
115	116	0.24567
116	117	0.24536
117	118	0.24525
118	119	0.24407
119	120	0.24386
120	121	0.24375
121	122	0.24365
122	123	0.24354
123	124	0.24322
124	125	0.24311
125	126	0.24279
126	127	0.24257
127	128	0.24247
128	129	0.24192
129	130	0.24148
130	131	0.24115
131	132	0.24093
132	133	0.24071
133	134	0.24060
134	135	0.24016

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 1 continued. Gamma Values for Population Cluster Structures

Population Gamma Index
2 to 262 Cluster Solutions

OBS	NCLUSTER	GAMMA
135	136	0.24005
136	137	0.23994
137	138	0.23983
138	139	0.23938
139	140	0.23927
140	141	0.23916
141	142	0.23904
142	143	0.23791
143	144	0.23769
144	145	0.23757
145	146	0.23712
146	147	0.23700
147	148	0.23689
148	149	0.23678
149	150	0.23655
150	151	0.23643
151	152	0.23631
152	153	0.23620
153	154	0.23608
154	155	0.23597
155	156	0.23585
156	157	0.23574
157	158	0.23550
158	159	0.23539
159	160	0.23515
160	161	0.23504
161	162	0.23469
162	163	0.23446
163	164	0.23434
164	165	0.23422
165	166	0.23305
166	167	0.23200
167	168	0.23105
168	169	0.23023
169	170	0.22952
170	171	0.22892
171	172	0.22844
172	173	0.22809
173	174	0.22785
174	175	0.22773
175	176	0.22713
176	177	0.22677
177	178	0.22629
178	179	0.22605
179	180	0.22569
180	181	0.22545
181	182	0.22532
182	183	0.22520
183	184	0.22508
184	185	0.22496
185	186	0.22387
186	187	0.22290
187	188	0.22253
188	189	0.22168
189	190	0.22094
190	191	0.22033
191	192	0.21983
192	193	0.21472
193	194	0.21434
194	195	0.21409
195	196	0.21396
196	197	0.21384
197	198	0.21371
198	199	0.21358
199	200	0.21346
200	201	0.21308
201	202	0.21282

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 1 continued. Gamma Values for Population Cluster Structures

Population Gamma Index
2 to 262 Cluster Solutions

OBS	NCLUSTER	GAMMA
202	203	0.21269
203	204	0.20755
204	205	0.20243
205	206	0.19732
206	207	0.19222
207	208	0.18714
208	209	0.18208
209	210	0.17703
210	211	0.17201
211	212	0.16699
212	213	0.16668
213	214	0.16167
214	215	0.15669
215	216	0.15601
216	217	0.15533
217	218	0.15482
218	219	0.15448
219	220	0.15397
220	221	0.15380
221	222	0.14874
222	223	0.14370
223	224	0.13867
224	225	0.13366
225	226	0.12866
226	227	0.12825
227	228	0.12326
228	229	0.11828
229	230	0.11333
230	231	0.10839
231	232	0.10347
232	233	0.10322
233	234	0.10272
234	235	0.10246
235	236	0.09752
236	237	0.09261
237	238	0.08771
238	239	0.08285
239	240	0.07802
240	241	0.07323
241	242	0.06849
242	243	0.06380
243	244	0.05919
244	245	0.05465
245	246	0.054179
246	247	0.049712
247	248	0.045375
248	249	0.041210
249	250	0.037273
250	251	0.033642
251	252	0.030429
252	253	0.027776
253	254	0.025858
254	255	0.024843
255	256	0.023785
256	257	0.021514
257	258	0.020283
258	259	0.018973
259	260	0.017565
260	261	0.012420
261	262	0.007170

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 2. Rand Analysis for Population

Rand Values Comparing 5 and 6-Cluster Structures
With 4 to 8-cluster Structures

	Number of Clusters				
	4	5	6	7	8
Number of Clusters					
5	0.9288	1.0000	0.9500	0.7832	0.7524
6	0.8793	0.9500	1.0000	0.8313	0.7999

Table 3. Comparison of 5- and 6-Cluster Structure in the Population

Rand Contingency Table Comparing
6-Cluster and 5-Cluster Solutions
CORRECTED RAND = 0.950020

	5-Cluster Solution					Row Totals
	1	2	3	4	5	
6-Cluster Solution						
1	41	0	0	0	0	41
2	17	0	0	0	0	17
3	0	44	0	0	0	44
4	0	0	119	0	0	119
5	0	0	0	18	0	18
6	0	0	0	0	24	24
Column Totals	58	44	119	18	24	263

APPENDIX D

ARMY POPULATION CLUSTER RESULTS

Table 4. Descriptive Analysis of Population 5-Cluster Structure

Ward Hierarchical Cluster Analysis
Five-Cluster Structure Solution

Cluster Name	N	Mean Factor Scores: FACTOR1 Ability	FACTOR2 Dexterity	FACTOR3 Outside	FACTOR4 Clerical	Distance Statistics: MIN	MAX	MEAN	STD
(1) Unskilled & Combat	58	-1.2387	-0.7288	-0.0491	-0.0679	0.4421	7.1720	2.9785	1.6027
(2) Mechanical	44	0.4210	0.2013	1.6060	0.3669	0.0668	5.9988	1.4692	1.1770
(3) Electrical Repair	119	0.2770	0.5562	-0.3569	-0.6087	0.0867	7.1285	1.0453	1.1128
(4) Clerical	18	-0.6321	0.8257	-0.8847	2.3896	0.1517	3.5774	1.2209	1.0397
(5) Technical	24	1.3225	-1.9848	-0.3927	0.7175	0.1460	3.1852	0.8726	0.7278

Table 5. Descriptive Analysis of Population 6-Cluster Structure

Ward Hierarchical Cluster Analysis
Six-Cluster Structure Solution

Cluster Name	N	Mean Factor Scores: FACTOR1 Ability	FACTOR2 Dexterity	FACTOR3 Outside	FACTOR4 Clerical	Distance Statistics: MIN	MAX	MEAN	STD
(1) Unskilled	41	-0.8579	-0.8216	-0.6515	0.0789	0.1184	4.3110	1.8513	1.3465
(2) Combat	17	-2.1570	-0.5051	1.4036	-0.4220	0.2554	4.0445	1.2702	0.9759
(3) Mechanical	44	0.4210	0.2013	1.6060	0.3669	0.0668	5.9988	1.4692	1.1770
(4) Electrical Repair	119	0.2770	0.5562	-0.3569	-0.6087	0.0867	7.1285	1.0453	1.1128
(5) Clerical	18	-0.6321	0.8257	-0.8847	2.3896	0.1517	3.5774	1.2209	1.0397
(6) Technical	24	1.3225	-1.9848	-0.3927	0.7175	0.1460	3.1852	0.8726	0.7278

APPENDIX E

SUGGESTED APPLICATIONS OF THE CV*IV PROCEDURE FOR DETERMINING THE POPULATION CLUSTER STRUCTURE

APPENDIX E

The CV*IV procedure was developed as a statistical approach for identifying the optimal number and configuration of clusters from a range of structures. We designed the procedure to test the following set of hypotheses:

H_o: the population is randomly distributed (i.e., there is no cluster structure in the population);

H_a: the population contains between 2 and n-1 clusters, where n is the number of objects in cross-samples.

Another possible application of the CV*IV procedure is to explore the statistical significance of a specific cluster structure, rather than a range of structures with different numbers of clusters. We do not recommend this application, but suspect that people will use it this way anyway. Therefore, we tentatively suggest a set of hypotheses, and caution the user that the results will not be conclusive about cluster structure.

H_o: the population is randomly distributed;

H_a: the population contains k clusters.

APPENDIX E

Example 1[1]

In this situation the user has measurements on a set of variables for a sample of jobs. The user thinks that there is a fairly strong job family structure in the population from which the sample was drawn--in this case, Army entry-level jobs. He or she examines structures with 2 to 10 clusters in the sample data. After reviewing the output of the CV*IV procedure, which follows, the user correctly rejects the null hypothesis at $p = .05$, and estimates that the population of entry-level jobs consists of 5 job families.

[1] Parts of this example are presented in the method section of the report. The full output from the CV*IV procedure is presented here.

APPENDIX E

Example 1

Summary of Cluster Structure Evaluation 6244

Observed Gamma Indices

NCLUSTER	OBS_GAM
2	0.290991
3	0.437476
4	0.366399
5	0.507991
6	0.376139
7	0.363164
8	0.394496
9	0.359222
10	0.359222

Summary of the Best Cluster Solution

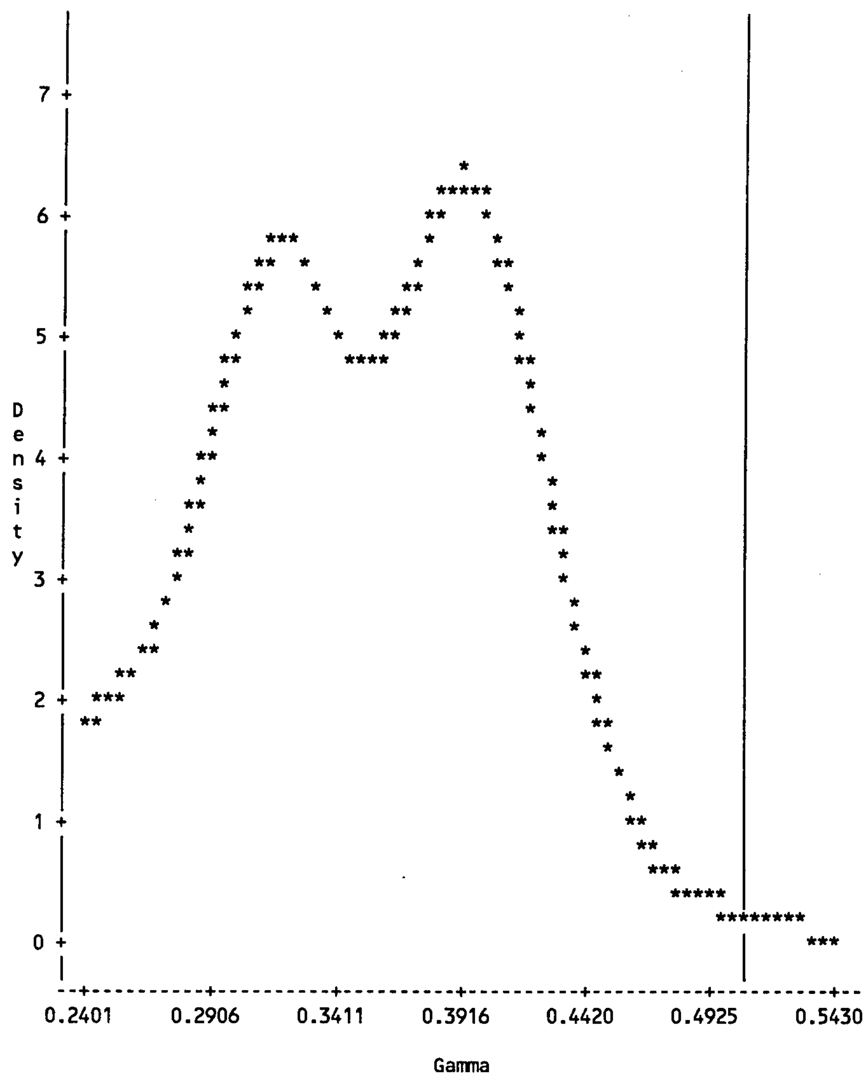
No. of Clusters	5
Observed Gamma	0.5079911
MC Mean of Gamma	0.352842
MC S.D. of Gamma	0.0035932
P-Value	0.01
No. of MC Samples	100

APPENDIX E

Example 1

```
                    Distribution of Gamma                          6245
        Under the Null Hypothesis of No Cluster Structure
                 Based on NORMAL Distribution

                  Structure =  5-Cluster Solution
            Gamma Index = 0.50799 (Reference Line)
                 Approximate P-value = 0.0100
```



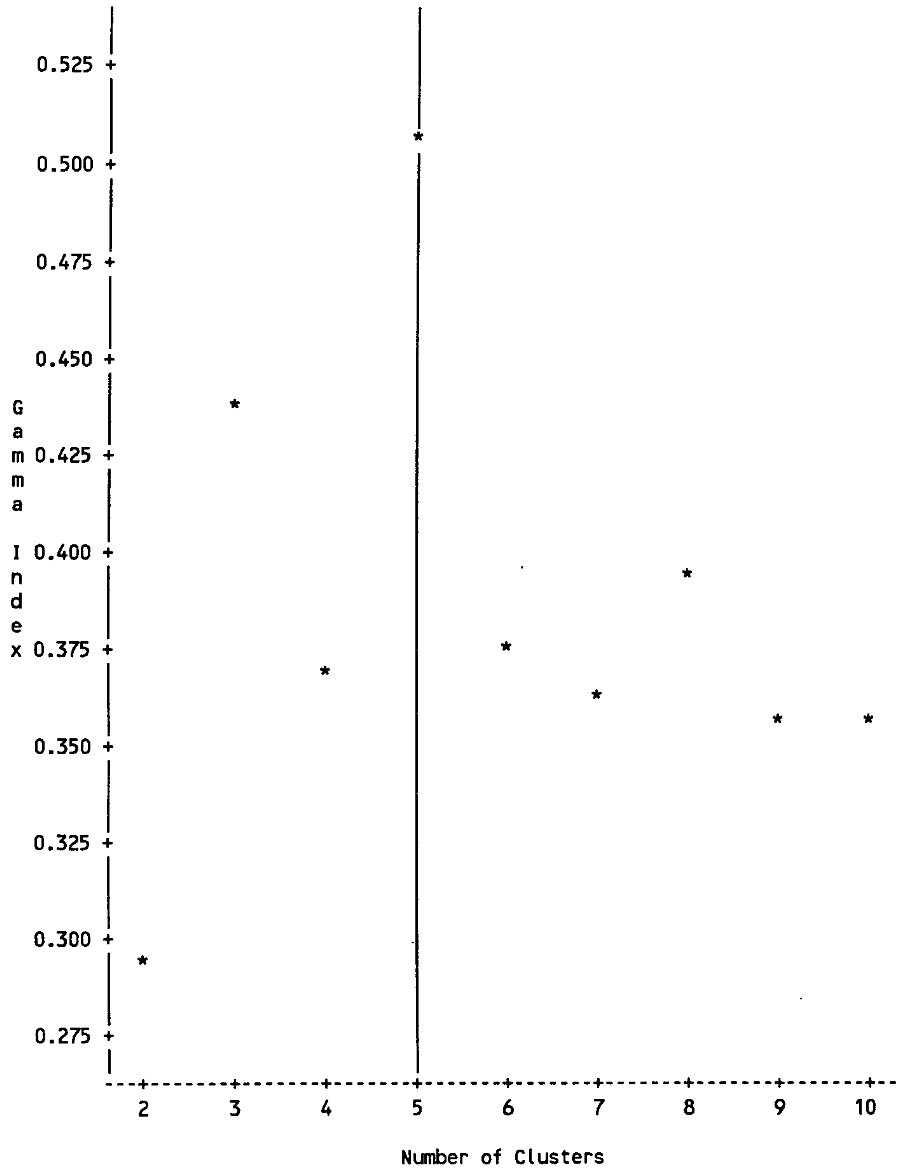
NOTE: 88 obs hidden.

APPENDIX E

Example 1

Summary of Cluster Structure Evaluation 6246
Plot of (Observed Gamma)*(Number of Clusters)

Plot of OBS_GAM*NCLUSTER. Symbol used is '*'.



APPENDIX E

Example 1

Optimal Cluster Structure Solution 6247

Cluster Assignments and Distances

-------------------------------- CLUSTER=1 --------------------------------

MOSID	DISTANCE
21L	0.11386
24C	0.11386
24K	0.11386
27F	0.11386
27L	0.11386
27N	0.11386
32F	0.11386
34C	0.11386
34E	0.11386
34Y	0.11386
31T	0.29831
35R	0.46523
63N	0.80893
84C	1.12621
35E	1.14585

-------------------------------- CLUSTER=2 --------------------------------

MOSID	DISTANCE
51R	0.23061
45B	0.79938
45K	0.79938
45T	0.79938
55G	0.79938
67G	1.12463
67H	1.12463
67U	1.12463
36E	1.19798
51C	2.85919

-------------------------------- CLUSTER=3 --------------------------------

MOSID	DISTANCE
11C	0.19004
16L	0.46830
16R	0.46830
62J	0.63830
19K	0.69898
11M	1.09386
17C	1.15058
62F	2.10312

-------------------------------- CLUSTER=4 --------------------------------

MOSID	DISTANCE
76C	0.71397

APPENDIX E

Example 1

Optimal Cluster Structure Solution 6248

Cluster Assignments and Distances

-------------------------------- CLUSTER=4 ---------------------------------
(continued)

MOSID	DISTANCE
76X	0.76297
76Y	0.87725
05D	1.26698
71P	2.67434
68F	3.16077
32H	4.15414
92B	8.54363

-------------------------------- CLUSTER=5 ---------------------------------

MOSID	DISTANCE
36K	0.89547
72G	1.63455
91U	1.81626
35H	1.83033
31V	2.22653
81B	2.53332
43M	2.77807
81C	4.10158
73C	5.08092

APPENDIX E

Example 1

Optimal Cluster Structure Solution 6249

Cluster Distances Statistics

Analysis Variable : DISTANCE

CLUSTER	N Obs	Mean	Std Dev	Minimum	Maximum
1	15	0.3322079	0.3794097	0.1138590	1.1458550
2	10	1.0859209	0.6848772	0.2306102	2.8591913
3	8	0.8514357	0.5993642	0.1900418	2.1031212
4	8	2.7692547	2.6594696	0.7139686	8.5436260
5	9	2.5441148	1.3039073	0.8954734	5.0809203

APPENDIX E

Example 1

Optimal Cluster Structure Solution 6250

Cluster Mean Factor Scores

CLUSTER	FACTOR1	FACTOR2	FACTOR3	FACTOR4
1	0.36445	-0.65706	0.86630	0.27260
2	1.00576	1.18632	-0.06173	0.23063
3	-1.42539	1.26293	0.05950	-0.01341
4	0.16579	-0.65183	-0.61309	-1.75758
5	-0.60528	-0.76624	-0.88315	0.86363

APPENDIX E

Example 1

Rand Contingency Table Comparing
5-Cluster and 4-Cluster Solutions
CORRECTED RAND = 0.712794

6251

	4-Cluster Solution				Row Totals
	CL1	CL2	CL3	CL4	
5-Cluster Solution					
CL1	15	0	0	0	15
CL2	0	10	0	0	10
CL3	0	0	8	0	8
CL4	0	0	0	8	8
CL5	9	0	0	0	9
Column Totals	24	10	8	8	50

APPENDIX E

Example 1

Rand Contingency Table Comparing
5-Cluster and 6-Cluster Solutions
CORRECTED RAND = 0.960455

6252

5-Cluster Solution	6-Cluster Solution CL1	CL2	CL3	CL4	CL5	CL6	Row Totals
CL1	15	0	0	0	0	0	15
CL2	0	10	0	0	0	0	10
CL3	0	0	8	0	0	0	8
CL4	0	0	0	5	0	3	8
CL5	0	0	0	0	9	0	9
Column Totals	15	10	8	5	9	3	50

APPENDIX E

Example 2

A researcher has a sample of Army jobs and believes that there is a 9-job family structure in the population. He examines only a 9-cluster structure and cannot reject the null hypothesis of no cluster structure. Since we have examined the Army job population cluster structure we know that it contains 5 clusters. However, the nonsignificant test results lead the user to incorrectly conclude that there is no cluster structure because he did not explore a range of possible numbers of clusters. We caution the researcher that the finding of nonsignificant results is not conclusive for this type of analysis and that he should examine a range of solutions. The output from his analysis are presented below.

APPENDIX E

Example 2

Summary of Cluster Structure Evaluation 3593

Observed Gamma Indices

NCLUSTER	OBS_GAM
9	0.359222

Summary of the Best Cluster Solution

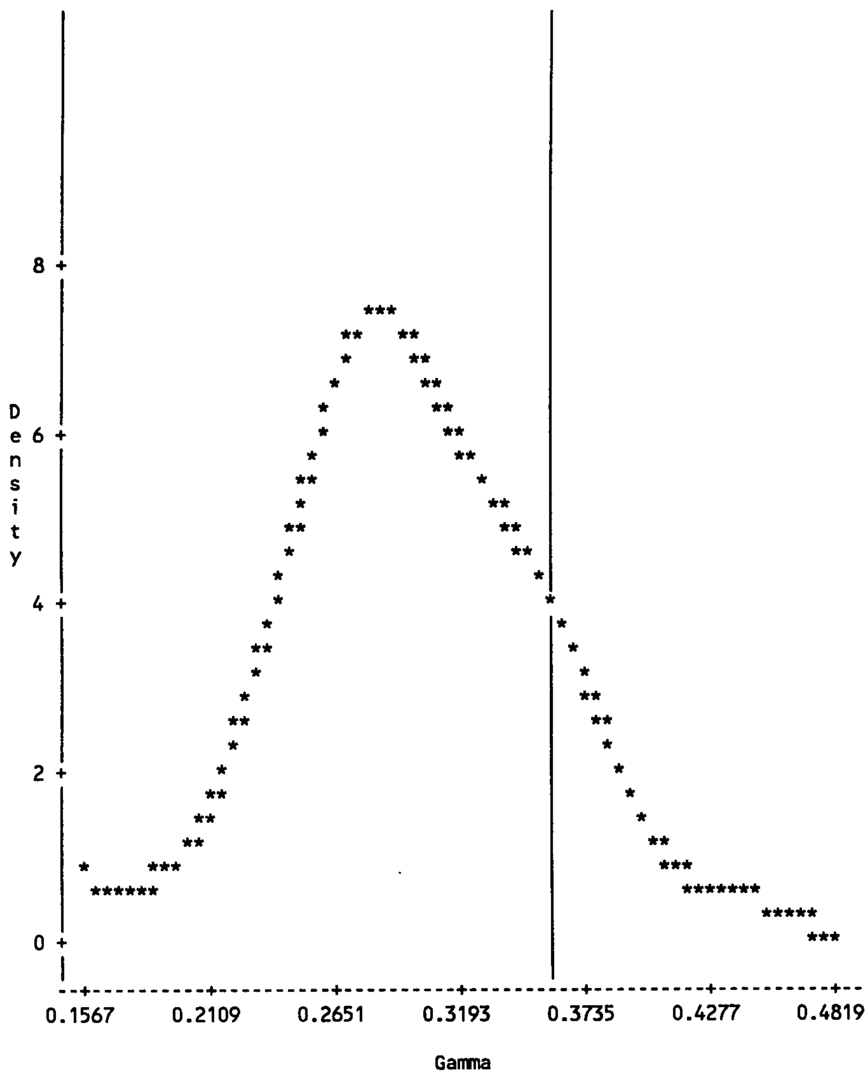
No. of Clusters	9
Observed Gamma	0.359222
MC Mean of Gamma	0.2991958
MC S.D. of Gamma	0.0034782
P-Value	0.17
No. of MC Samples	100

APPENDIX E

Example 2

Distribution of Gamma   3594
Under the Null Hypothesis of No Cluster Structure
Based on NORMAL Distribution

Structure = 9-Cluster Solution
Gamma Index = 0.35922 (Reference Line)
Approximate P-value = 0.1700



NOTE: 105 obs hidden.

APPENDIX E

Example 2

Optimal Cluster Structure Solution 3596

Cluster Assignments and Distances

--------------------------------- CLUSTER=1 ---------------------------------

MOSID	DISTANCE
21L	0.11386
24C	0.11386
24K	0.11386
27F	0.11386
27L	0.11386
27N	0.11386
32F	0.11386
34C	0.11386
34E	0.11386
34Y	0.11386
31T	0.29831
35R	0.46523
63N	0.80893
84C	1.12621
35E	1.14585

--------------------------------- CLUSTER=2 ---------------------------------

MOSID	DISTANCE
45B	0.04435
45K	0.04435
45T	0.04435
55G	0.04435
51R	0.38673
36E	0.74184

--------------------------------- CLUSTER=3 ---------------------------------

MOSID	DISTANCE
11C	0.19004
16L	0.46830
16R	0.46830
62J	0.63830
19K	0.69898
11M	1.09386
17C	1.15058
62F	2.10312

--------------------------------- CLUSTER=4 ---------------------------------

MOSID	DISTANCE
67G	0.10441
67H	0.10441
67U	0.10441
51C	0.93965

Example 2

Optimal Cluster Structure Solution 3597

Cluster Assignments and Distances

-------------------------------- CLUSTER=5 --------------------------------

MOSID	DISTANCE
76X	0.41498
76Y	0.62194
76C	0.69897
68F	0.98607
32H	1.48145

-------------------------------- CLUSTER=6 --------------------------------

MOSID	DISTANCE
36K	0.27115
35H	0.36787
43M	0.80942
91U	1.56327

-------------------------------- CLUSTER=7 --------------------------------

MOSID	DISTANCE
31V	0.25007
81B	0.54987
72G	1.04323

-------------------------------- CLUSTER=8 --------------------------------

MOSID	DISTANCE
71P	0.43692
05D	1.44326
92B	1.49530

-------------------------------- CLUSTER=9 --------------------------------

MOSID	DISTANCE
81C	1.26243
73C	1.26243

APPENDIX E

Example 2

Optimal Cluster Structure Solution 3598

Cluster Distances Statistics

Analysis Variable : DISTANCE

CLUSTER	N Obs	Mean	Std Dev	Minimum	Maximum
1	15	0.3322079	0.3794097	0.1138590	1.1458550
2	6	0.2176645	0.2910283	0.0443547	0.7418398
3	8	0.8514357	0.5993642	0.1900418	2.1031212
4	4	0.3132172	0.4176230	0.1044057	0.9396517
5	5	0.8406816	0.4125763	0.4149810	1.4814525
6	4	0.7529285	0.5888440	0.2711541	1.5632662
7	3	0.6143903	0.4004971	0.2500714	1.0432311
8	3	1.1251615	0.5965982	0.4369245	1.4952951
9	2	1.2624350	0	1.2624350	1.2624350

APPENDIX E

Example 2

Optimal Cluster Structure Solution 3599

Cluster Mean Factor Scores

CLUSTER	FACTOR1	FACTOR2	FACTOR3	FACTOR4
1	0.36445	-0.65706	0.86630	0.27260
2	1.21362	1.28417	-0.54242	0.74978
3	-1.42539	1.26293	0.05950	-0.01341
4	0.69397	1.03954	0.65930	-0.54809
5	-0.13371	-0.58895	0.37773	-1.89088
6	-1.42211	-0.79469	-0.42834	0.41394
7	0.46179	-0.34605	-0.98334	0.42563
8	0.66495	-0.75662	-2.26445	-1.53543
9	-0.57221	-1.33961	-1.64250	2.42003

APPENDIX E

Example 2

Rand Contingency Table Comparing
9-Cluster and 8-Cluster Solutions
CORRECTED RAND = 0.961462

3600

	8-Cluster Solution							
9-Cluster Solution	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8
CL1	15	0	0	0	0	0	0	0
CL2	0	6	0	0	0	0	0	0
CL3	0	0	8	0	0	0	0	0
CL4	0	0	0	4	0	0	0	0
CL5	0	0	0	0	5	0	0	0
CL6	0	0	0	0	0	4	0	0
CL7	0	0	0	0	0	3	0	0
CL8	0	0	0	0	0	0	3	0
CL9	0	0	0	0	0	0	0	2
Column Totals	15	6	8	4	5	7	3	2

(CONTINUED)

APPENDIX E

Example 2

Rand Contingency Table Comparing
9-Cluster and 8-Cluster Solutions
CORRECTED RAND = 0.961462

3601

	Row Totals
9-Cluster Solution	
CL1	15
CL2	6
CL3	8
CL4	4
CL5	5
CL6	4
CL7	3
CL8	3
CL9	2
Column Totals	50

APPENDIX E

Example 2

Rand Contingency Table Comparing 3602
9-Cluster and 10-Cluster Solutions
CORRECTED RAND = 0.979905

	10-Cluster Solution							
9-Cluster Solution	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8
CL1	15	0	0	0	0	0	0	0
CL2	0	6	0	0	0	0	0	0
CL3	0	0	8	0	0	0	0	0
CL4	0	0	0	4	0	0	0	0
CL5	0	0	0	0	3	0	2	0
CL6	0	0	0	0	0	4	0	0
CL7	0	0	0	0	0	0	0	3
CL8	0	0	0	0	0	0	0	0
CL9	0	0	0	0	0	0	0	0
Column Totals	15	6	8	4	3	4	2	3

(CONTINUED)

APPENDIX E

Example 2

Rand Contingency Table Comparing
9-Cluster and 10-Cluster Solutions
CORRECTED RAND = 0.979905

3603

	10-Cluster Solution		
	CL9	CL10	Row Totals
9-Cluster Solution			
CL1	0	0	15
CL2	0	0	6
CL3	0	0	8
CL4	0	0	4
CL5	0	0	5
CL6	0	0	4
CL7	0	0	3
CL8	3	0	3
CL9	0	2	2
Column Totals	3	2	50

APPENDIX E

Example 3

A different user has the same sample of Army jobs. However, she believes that there is a 3-job family structure in the population. She examines only a 3-cluster structure and is able to reject the null hypothesis. She concludes that the population consists of 3 job families. Again, we have examined the Army population cluster structure and know that a 5- or 6-cluster structure has a higher level of internal validity. We caution the user to examine a range of cluster structures because the 3-cluster structure may not be optimal. The output from the 3-cluster analysis is presented below.

APPENDIX E

Example 3

Summary of Cluster Structure Evaluation 405

Observed Gamma Indices

NCLUSTER	OBS_GAM
3	0.437476

Summary of the Best Cluster Solution

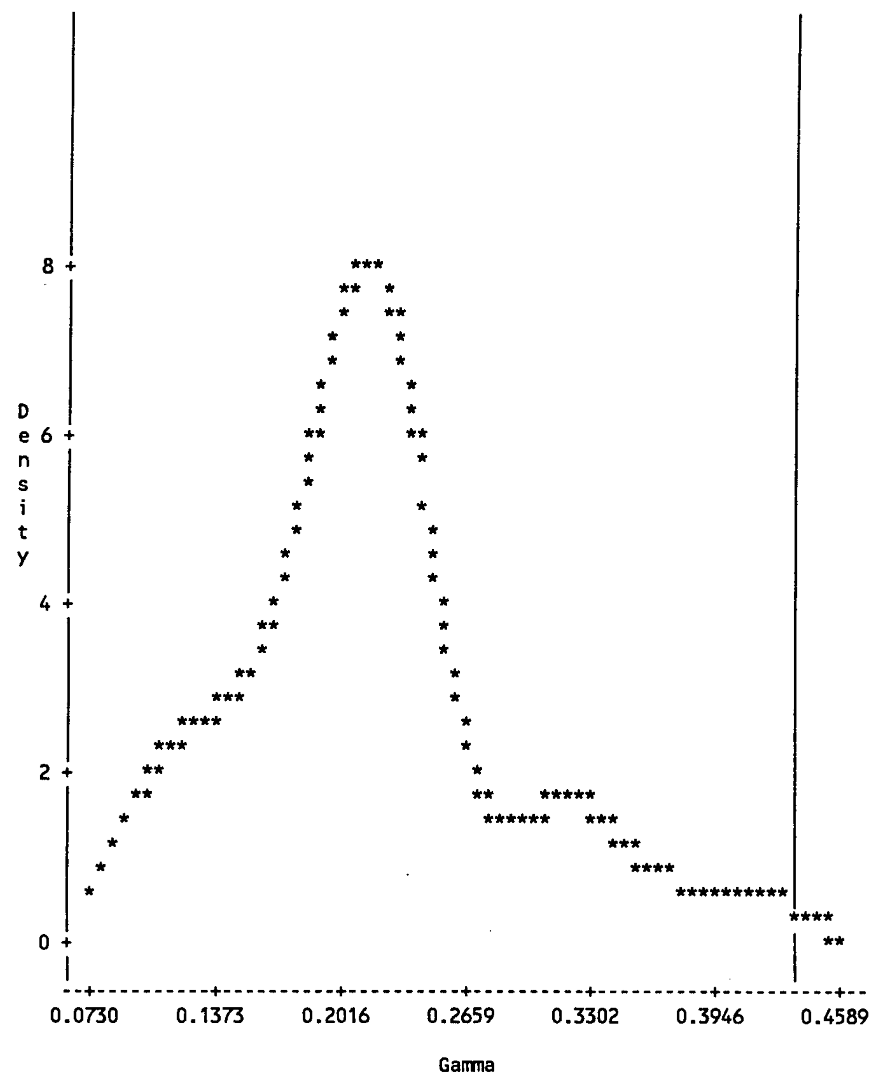
No. of Clusters	3
Observed Gamma	0.4374765
MC Mean of Gamma	0.2183129
MC S.D. of Gamma	0.0050616
P-Value	0
No. of MC Samples	100

APPENDIX E

Example 3

```
                 Distribution of Gamma                              406
      Under the Null Hypothesis of No Cluster Structure
                Based on NORMAL Distribution

               Structure =  3-Cluster Solution
            Gamma Index = 0.43748 (Reference Line)
                Approximate P-value = 0.0000
```



NOTE: 97 obs hidden.

APPENDIX E

Example 3

Optimal Cluster Structure Solution 408

Cluster Assignments and Distances

-------------------------------- CLUSTER=1 ---------------------------------

MOSID	DISTANCE
31T	0.5823
21L	0.7961
24C	0.7961
24K	0.7961
27F	0.7961
27L	0.7961
27N	0.7961
32F	0.7961
34C	0.7961
34E	0.7961
34Y	0.7961
35R	1.2123
81B	1.2637
31V	1.3791
63N	1.3940
84C	1.4561
35E	1.5789
35H	2.5488
43M	2.8247
36K	3.4243
91U	4.0070
72G	4.4977
81C	6.9344
73C	11.3532

-------------------------------- CLUSTER=2 ---------------------------------

MOSID	DISTANCE
62F	0.13055
62J	0.99867
67G	1.58579
67H	1.58579
67U	1.58579
36E	1.70396
51R	2.27832
11C	2.42504
16L	2.50711
16R	2.50711
51C	2.56740
45B	2.84637
45K	2.84637
45T	2.84637
55G	2.84637
19K	2.85724
11M	4.74816
17C	5.42926

APPENDIX E

Example 3

Optimal Cluster Structure Solution 409

Cluster Assignments and Distances

-------------------------------- CLUSTER=3 ---------------------------------

MOSID	DISTANCE
76C	0.71397
76X	0.76297
76Y	0.87725
05D	1.26698
71P	2.67434
68F	3.16077
32H	4.15414
92B	8.54363

APPENDIX E

Example 3

Optimal Cluster Structure Solution 410

Cluster Distances Statistics

Analysis Variable : DISTANCE

CLUSTER	N Obs	Mean	Std Dev	Minimum	Maximum
1	24	2.1840619	2.4950970	0.5822800	11.3532367
2	18	2.4608705	1.2195333	0.1305536	5.4292589
3	8	2.7692547	2.6594696	0.7139686	8.5436260

APPENDIX E

Example 3

Optimal Cluster Structure Solution 411

Cluster Mean Factor Scores

CLUSTER	FACTOR1	FACTOR2	FACTOR3	FACTOR4
1	0.00080	-0.69800	0.21025	0.49423
2	-0.07475	1.22037	-0.00785	0.12217
3	0.16579	-0.65183	-0.61309	-1.75758

APPENDIX E

Example 3

Rand Contingency Table Comparing
3-Cluster and 2-Cluster Solutions
CORRECTED RAND = 0.691202

	2-Cluster Solution		Row Totals
	CL1	CL2	
3-Cluster Solution			
CL1	24	0	24
CL2	0	18	18
CL3	8	0	8
Column Totals	32	18	50

APPENDIX E

Example 3

Rand Contingency Table Comparing
3-Cluster and 4-Cluster Solutions
CORRECTED RAND = 0.855259

413

	4-Cluster Solution				Row Totals
	CL1	CL2	CL3	CL4	
3-Cluster Solution					
CL1	24	0	0	0	24
CL2	0	10	8	0	18
CL3	0	0	0	8	8
Column Totals	24	10	8	8	50

APPENDIX E

Example 4

In this analysis a researcher believes that Army recruits can be grouped into distinct clusters according to profiles of the 10 tests of the Armed Forces Vocational Aptitude Battery (ASVAB). She selects a sample of new recruits and examines structures with 2 to 20 clusters. The results of the CV*IV procedure shown below are not significant. Therefore, she cannot reject the null hypothesis that the population distribution of recruits is approximately multivariate random normal.

APPENDIX E

Example 4

Summary of Cluster Structure Evaluation 12403

Observed Gamma Indices

NCLUSTER	OBS_GAM
2	0.139594
3	0.127699
4	0.163875
5	0.113496
6	0.131078
7	0.233124
8	0.209863
9	0.193008
10	0.180168
11	0.190933
12	0.225885
13	0.217925
14	0.260830
15	0.260418
16	0.269178
17	0.267467
18	0.267467
19	0.265964
20	0.229481

Summary of the Best Cluster Solution

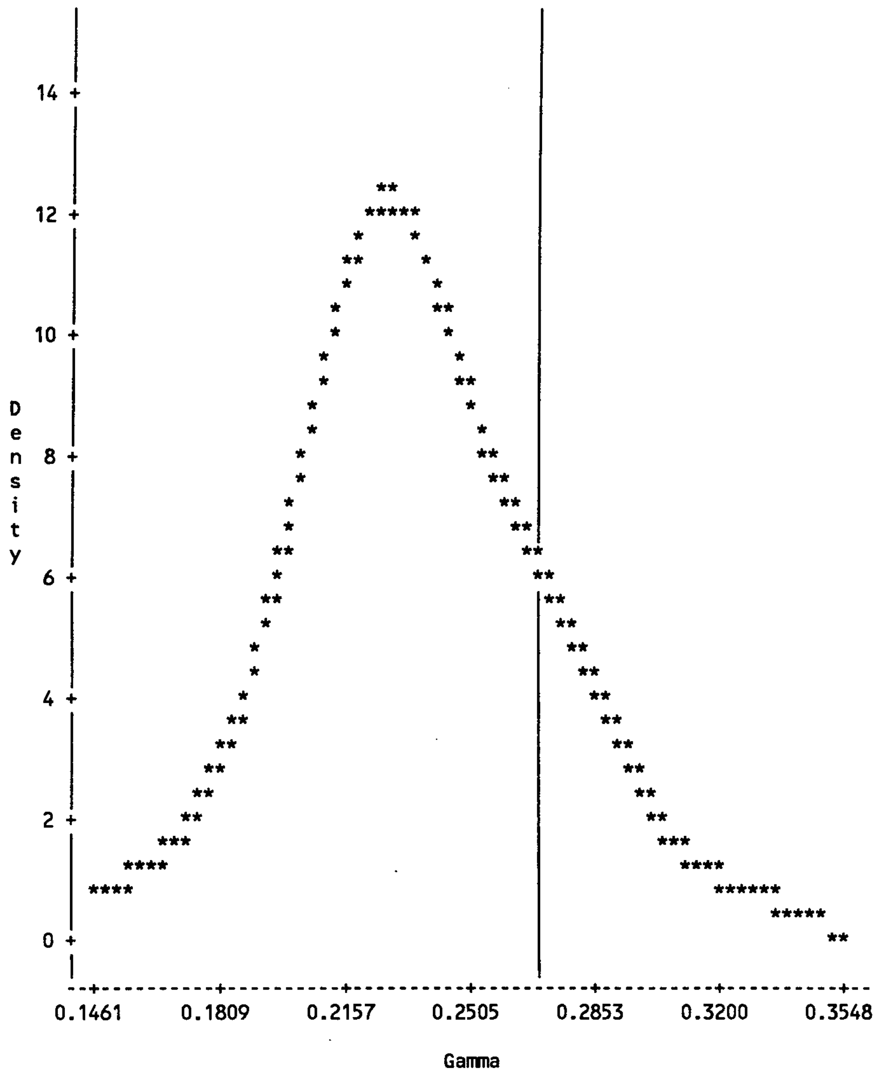
No. of Clusters	16
Observed Gamma	0.2691784
MC Mean of Gamma	0.2354532
MC S.D. of Gamma	0.0013764
P-Value	0.18
No. of MC Samples	100

APPENDIX E

Example 4

Distribution of Gamma 12404
Under the Null Hypothesis of No Cluster Structure
Based on NORMAL Distribution

Structure = 16-Cluster Solution
Gamma Index = 0.26918 (Reference Line)
Approximate P-value = 0.1800



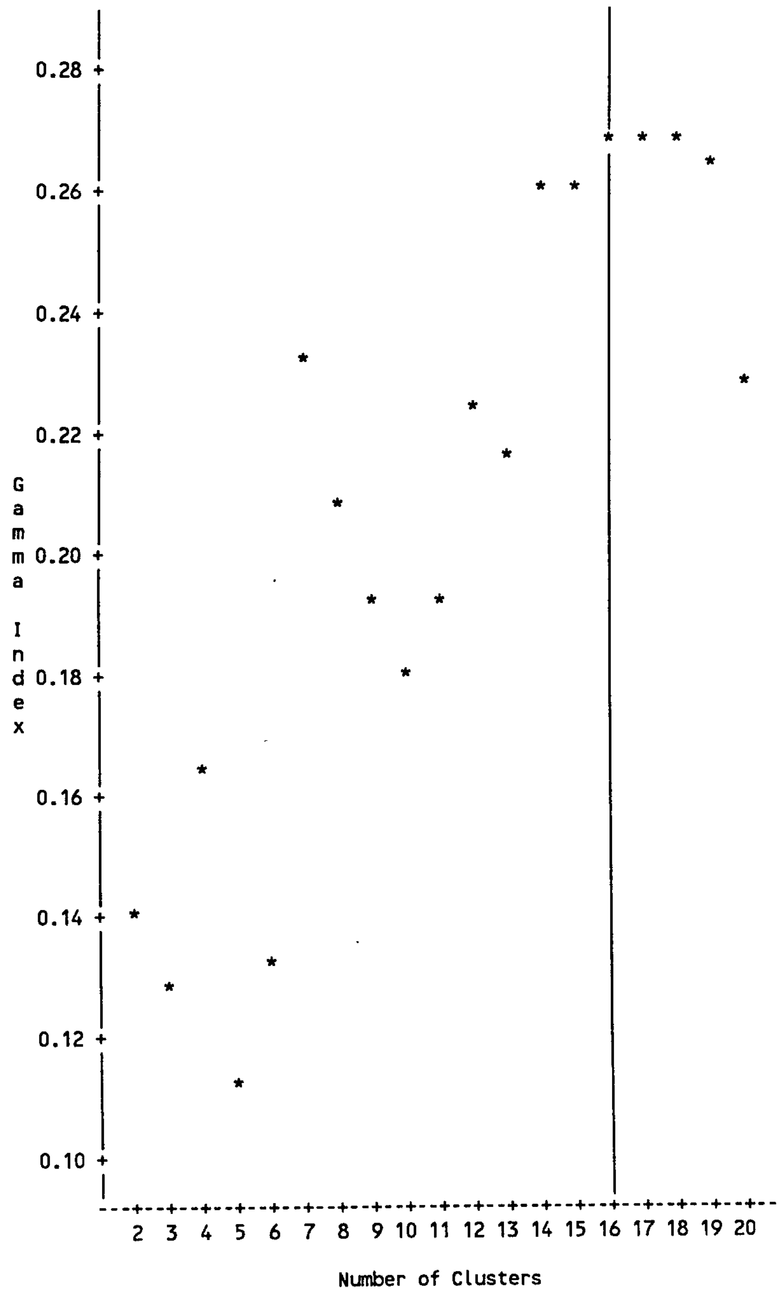
NOTE: 86 obs hidden.

APPENDIX E

Example 4

Summary of Cluster Structure Evaluation 12405
Plot of (Observed Gamma)*(Number of Clusters)

Plot of OBS_GAM*NCLUSTER. Symbol used is '*'.



APPENDIX E

Example 4

Optimal Cluster Structure Solution 12409

Cluster Distances Statistics

Analysis Variable : DISTANCE

CLUSTER	N Obs	Mean	Std Dev	Minimum	Maximum
1	8	2.1791234	0.8381630	1.2688995	3.9377852
2	4	2.4062900	0.2784894	2.0333227	2.6669408
3	5	3.7097727	0.9240487	2.7209232	5.1504901
4	3	1.5544387	0.7935117	0.6961742	2.2614150
5	6	3.5694026	1.2716619	1.9380023	5.3368007
6	5	3.8074435	1.0992187	2.6720207	5.4311877
7	8	5.2434283	2.8788394	2.8393306	10.7832645
8	4	2.7990472	1.0295825	1.4579630	3.6852831
9	3	2.3620342	0.8325038	1.6011149	3.2512314
10	6	4.0341251	1.2375640	2.9104776	6.0515814
11	4	5.0733972	1.9020194	2.9255975	7.4480957
12	4	4.4499681	2.3481337	2.2627003	6.6972429
13	4	4.4847027	0.7691011	3.7403477	5.5633355
14	2	2.4499192	0	2.4499192	2.4499192
15	2	2.6216764	0	2.6216764	2.6216764
16	2	2.7477508	0	2.7477508	2.7477508

APPENDIX E

Example 4

Optimal Cluster Structure Solution 12410

Cluster Mean Factor Scores

CLUSTER	FACTOR1	FACTOR2	FACTOR3	FACTOR4
1	-0.06057	0.69340	0.84886	0.40579
2	0.57819	-0.76211	-0.07276	-1.42507
3	-0.62543	-0.85683	-0.31171	-0.61986
4	0.09420	1.16196	0.11421	0.53987
5	-0.87416	0.54167	-0.43775	-0.87051
6	-0.16333	-1.19708	-0.15252	0.88440
7	0.48668	-0.38458	1.15709	-0.20058
8	-0.33736	0.55424	0.58518	1.52224
9	0.85790	1.41373	-0.79125	-1.14721
10	0.63570	-0.90590	-0.57684	0.19740
11	-1.14684	0.26740	-1.41594	0.27333
12	-0.22808	-0.03345	-0.59009	0.62282
13	1.01014	0.92549	0.47648	-0.79621
14	0.40448	-0.68914	-0.43389	0.94035
15	-1.35678	0.97114	0.76312	0.19369
16	0.75491	-1.05650	-1.09886	-0.08012

FACTOR5	FACTOR6	FACTOR7	FACTOR8	FACTOR9
0.48751	0.62059	0.22526	0.11317	0.78482
0.62854	0.60419	0.44139	0.44706	0.70286
-0.14371	-1.25318	-0.95666	0.50817	0.11001
-1.15763	-0.48265	-0.57938	-1.41932	-0.21590
-0.03931	0.97509	-0.04304	-0.64739	0.18584
-0.06666	-1.12291	0.21119	-1.18717	1.22399
-0.87931	0.33225	0.90795	0.10771	-0.79646
0.58841	0.45539	-0.56628	1.24269	-0.44470
0.62870	-0.37246	0.33664	-0.66786	0.75513
0.21299	0.80669	-0.94436	0.98093	-0.01419
-1.82393	0.87836	-0.06494	0.17799	0.13663
1.29639	-0.54586	1.28630	0.68459	0.18554
-0.78633	-1.81885	-0.62112	0.57383	-0.39387
1.26680	0.72033	-1.04306	-2.07308	-1.54004
1.13643	-1.53771	-0.63449	-0.67281	-1.13225
0.15608	-0.26287	1.38400	-0.56227	-2.31285

APPENDIX E

Example 4

Rand Contingency Table Comparing
16-Cluster and 15-Cluster Solutions
CORRECTED RAND = 0.978407

12411

16-Cluster Solution	15-Cluster Solution CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8
CL1	8	0	0	0	0	0	0	0
CL2	0	4	0	0	0	0	0	0
CL3	0	0	5	0	0	0	0	0
CL4	0	0	0	3	0	0	0	0
CL5	0	0	0	0	6	0	0	0
CL6	0	0	0	0	0	5	0	0
CL7	0	0	0	0	0	0	8	0
CL8	0	0	0	0	0	0	0	4
CL9	0	0	0	0	0	0	0	0
CL10	0	0	0	0	0	0	0	0
CL11	0	0	0	0	0	0	0	0
CL12	0	0	0	0	0	0	0	0
CL13	0	0	0	0	0	0	0	0
CL14	0	0	0	0	0	0	0	0
CL15	0	0	0	2	0	0	0	0
CL16	0	0	0	0	0	0	0	0
Column Totals	8	4	5	5	6	5	8	4

(CONTINUED)

APPENDIX E

Example 4

Rand Contingency Table Comparing
16-Cluster and 15-Cluster Solutions
CORRECTED RAND = 0.978407

12412

	15-Cluster Solution							Row Totals
	CL9	CL10	CL11	CL12	CL13	CL14	CL15	
16-Cluster Solution								
CL1	0	0	0	0	0	0	0	8
CL2	0	0	0	0	0	0	0	4
CL3	0	0	0	0	0	0	0	5
CL4	0	0	0	0	0	0	0	3
CL5	0	0	0	0	0	0	0	6
CL6	0	0	0	0	0	0	0	5
CL7	0	0	0	0	0	0	0	8
CL8	0	0	0	0	0	0	0	4
CL9	3	0	0	0	0	0	0	3
CL10	0	6	0	0	0	0	0	6
CL11	0	0	4	0	0	0	0	4
CL12	0	0	0	4	0	0	0	4
CL13	0	0	0	0	4	0	0	4
CL14	0	0	0	0	0	2	0	2
CL15	0	0	0	0	0	0	0	2
CL16	0	0	0	0	0	0	2	2
Column Totals	3	6	4	4	4	2	2	70

APPENDIX E

Example 4

Rand Contingency Table Comparing
16-Cluster and 17-Cluster Solutions
CORRECTED RAND = 0.973727

12413

16-Cluster Solution	17-Cluster Solution CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8
CL1	8	0	0	0	0	0	0	0
CL2	0	4	0	0	0	0	0	0
CL3	0	0	5	0	0	0	0	0
CL4	0	0	0	3	0	0	0	0
CL5	0	0	0	0	6	0	0	0
CL6	0	0	0	0	0	5	0	0
CL7	0	0	0	0	0	0	7	0
CL8	0	0	0	0	0	0	0	4
CL9	0	0	0	0	0	0	0	0
CL10	0	0	0	0	0	0	0	0
CL11	0	0	0	0	0	0	0	0
CL12	0	0	0	0	0	0	0	0
CL13	0	0	0	0	0	0	0	0
CL14	0	0	0	0	0	0	0	0
CL15	0	0	0	0	0	0	0	0
CL16	0	0	0	0	0	0	0	0
Column Totals	8	4	5	3	6	5	7	4

(CONTINUED)

APPENDIX E

Example 4

Rand Contingency Table Comparing 12414
16-Cluster and 17-Cluster Solutions
CORRECTED RAND = 0.973727

16-Cluster Solution	17-Cluster Solution							
	CL9	CL10	CL11	CL12	CL13	CL14	CL15	CL16
CL1	0	0	0	0	0	0	0	0
CL2	0	0	0	0	0	0	0	0
CL3	0	0	0	0	0	0	0	0
CL4	0	0	0	0	0	0	0	0
CL5	0	0	0	0	0	0	0	0
CL6	0	0	0	0	0	0	0	0
CL7	0	0	0	0	0	0	0	0
CL8	0	0	0	0	0	0	0	0
CL9	3	0	0	0	0	0	0	0
CL10	0	6	0	0	0	0	0	0
CL11	0	0	4	0	0	0	0	0
CL12	0	0	0	4	0	0	0	0
CL13	0	0	0	0	4	0	0	0
CL14	0	0	0	0	0	2	0	0
CL15	0	0	0	0	0	0	2	0
CL16	0	0	0	0	0	0	0	2
Column Totals	3	6	4	4	4	2	2	2

(CONTINUED)

APPENDIX E

Example 4

Rand Contingency Table Comparing
16-Cluster and 17-Cluster Solutions
CORRECTED RAND = 0.973727

12415

	17-Cluster Solution CL17	Row Totals
16-Cluster Solution		
CL1	0	8
CL2	0	4
CL3	0	5
CL4	0	3
CL5	0	6
CL6	0	5
CL7	1	8
CL8	0	4
CL9	0	3
CL10	0	6
CL11	0	4
CL12	0	4
CL13	0	4
CL14	0	2
CL15	0	2
CL16	0	2
Column Totals	1	70